PROGRAMME AND ABSTRACTS

CFE-CMStatistics 2024

18th International Conference on

Computational and Financial Econometrics (CFE 2024)

and

Computational and Methodological Statistics (CMStatistics 2024)

https://www.cmstatistics.org/CFECMStatistics2024

King's College London, UK

14 – 16 December 2024

ISBN 978-9925-7812-8-7

Co-chairs:
Mario Peruggia, Tommaso Proietti, Matthieu Marbac, Willi Semmler and Svetlana Makarova.

International Organizing Committee:
Ana Colubi, Erricos Kontoghiorghes and Michael Pitt.

Scientific Programme Committee:

Alessandra Amendola, Josu Arteche, Andreas Artemiou, Eric Beutner, Monica Billio, Enea Bongiorno, Federico Camerlenghi, Massimo Cannas, Joshua Cape, Stefano Castruccio, Yoosoon Chang, Cathy W.S. Chen, Eliana Christou, Abdelaati Daouia, Liqun Diao, Antoine Djogbenou, Marie Du Roy, Takeshi Emura, M. Brigida Ferraro, Frederic Ferraty, Marina Friedrich, Jairo Fuquene, Jan Gertheiss, Subharup Guha, Rajarshi Guhaniyogi, Marc Hallin, Tony He, Alain Hecq, Masayuki Hirukawa, Galin Jones, Menelaos Karanasos, Sayar Karmakar, Michail Karoglou, Deborah Kunkel, Andrew Lawson, Kuang-Yao Lee, Keith Levin, Cai Li, Degui Li, Meng Li, Tsung-I Lin, Nicola Loperfido, Francesco Simone Ludici, Shujie Ma, Ranjan Maitra, Simone Manganelli, Paolo Maranzano, Etienne Marceau, Hiroki Masuda, Lorenzo Mercuri, Michelle Miranda, Kathrin Moellenhoff, Pavlo Mozharovskyi, Kalliopi Mylona, Chris Otrock, Michael Owyang, Alessia Pini, Artem Prokhorov, Monia Ranalli, Paulo Rodrigues, David Rossell, Vivekananda Roy, Javier Rubio, Joachim Schnurbus, Michael Schweinberger, Catia Scricciolo, Chengchun Shi, Anton Skrobotov, Mike So, John Stufken, Wei Sun, Vincent Vandewalle, Domenico Vitale, Weining Wang, Yuedong Wang, Toshiaki Watanabe, Marten Wegkamp, Andrew Wood, Stefan Wrzaczek, Nakahiro Yoshida, Qingzhao Yu, Yunpeng Zhao, Ping-Shou Zhong, Shuheng Zhou, Wen Zhou and Ines del Puerto.

Local Organizer:
King's Business School and King's Department of Mathematics.
CFEnetwork and CMStatistics.

Dear Friends and Colleagues,

We warmly welcome you to London for the 18th International Conference on Computational and Financial Econometrics (CFE 2024) and Computational and Methodological Statistics (CMStatistics 2024).

The conference aims to bring together researchers and practitioners to discuss recent methodology and computational approaches for economics, finance, and statistics. The CFE-CMStatistics 2024 programme consists of about 370 sessions, four plenary talks, and nearly 1500 presentations. With over 1600 participants, this conference stands out as one of the most important international scientific events in the field.

The co-chairs have endeavoured to provide a balanced and stimulating programme that will appeal to the diverse interests of the participants. The international organizing committee hopes that the hybrid conference will provide an ideal environment to communicate effectively with colleagues. The conference is the collective effort of many individuals and organizations. The Scientific Programme Committee, the Session Organizers, the supporting universities, and various agents have contributed substantially to the organization of the conference. We acknowledge their work and the support of our networks.

King's College London (KCL) offers excellent facilities and a fantastic environment in central London. Through their efforts, the local hosts and sponsoring organizations have substantially contributed to the successful organization of the conference. We thank them all for their support. In particular, we express our sincere appreciation to the hosts, the Department of Mathematics at KCL and the Data Analytics for Finance and Macro (DAFM) Research Centre at the King's Business School.

We are pleased to announce that the official journal of CFEnetwork and CMStatistics, EcoSta, has an impact factor of 2.0, ranking in high positions in the related areas. CMStatistics also publishes The Annals of Statistical Data Science (SDS) as a supplement to the Elsevier journal Computational Statistics & Data Analysis (CSDA). The CSDA is the official journal of CMStatistics as well. CSDA continues to uphold its commendable and consistent performance, with an impact factor of 1.5. You are encouraged to submit your papers to EcoSta, the Annals of SDS or regular peer-reviewed issues of CSDA.

Looking ahead, the CFE-CMStatistics 2025 conference will be hosted at King's College London, UK, from Saturday, December 13th, to Monday, December 15th, 2025, with tutorials scheduled prior to the conference. We extend a heartfelt invitation and enthusiastic encouragement for your active participation in these forthcoming events.

We wish you a productive and stimulating conference.

Kind regards,

Ana Colubi, Erricos J. Kontoghiorghes and Manfred Deistler
Coordinators of CMStatistics & CFEnetwork and EcoSta.

CMStatistics: ERCIM Working Group on COMPUTATIONAL AND METHODOLOGICAL STATISTICS

http://www.cmstatistics.org

The working group (WG) CMStatistics comprises a number of specialized teams in various research areas of computational and methodological statistics. The teams act autonomously within the framework of the WG in order to promote their own research agenda. Their activities are endorsed by the WG. They submit research proposals, organize sessions, tracks and tutorials during the annual WG meetings and edit journal special issues. The Econometrics and Statistics (EcoSta) and Computational Statistics & Data Analysis (CSDA) are the official journals of the CMStatistics.

Specialized teams

Currently, the ERCIM WG has over 1950 members and the following specialized teams

BIO:	Biostatistics
BS:	Bayesian Statistics
DMC:	Dependence Models and Copulas
DOE:	Design Of Experiments
FDA:	Functional Data Analysis
HDS:	High-Dimensional Statistics
IS:	Imprecision in Statistics
LVSEM:	Latent Variable and Structural Equation Models
MM:	Mixture Models
NPS:	Non-Parametric Statistics
RS:	Robust Statistics
SA:	Survival Analysis
SAE:	Small Area Estimation
SDS:	Statistical Data Science: Methods and Computations
SEA:	Statistics of Extremes and Applications
SL:	Statistical Learning
TSMC:	Times Series

You are encouraged to become a member of the WG. For further information, please contact the Chairs of the specialized groups (see the WG's website) or email at info@cmstatistics.org.

CFEnetwork COMPUTATIONAL AND FINANCIAL ECONOMETRICS

http://www.CFEnetwork.org

The Computational and Financial Econometrics (CFEnetwork) comprises a number of specialized teams in various research areas of theoretical and applied econometrics, financial econometrics and computation, and empirical finance. The teams contribute to the network's activities by organizing sessions, tracks and tutorials during the annual CFEnetwork meetings, and by submitting research proposals. Furthermore, the teams edit special issues currently published under the Annals of CFE. The Econometrics and Statistics (EcoSta) is the official journal of the CFEnetwork. Currently, the CFEnetwork has over 1100 members.

You are encouraged to become a member of the CFEnetwork. For further information, please see the website or contact by email at info@cfenetwork.org.

SCHEDULE (GMT)

2024-12-14

Opening, 08:35 - 08:45

A - Keynote
CFECMStatistics2024
08:45 - 09:35

Coffee break
09:35 - 10:05

B
CFECMStatistics2024
10:05 - 12:10

Lunch break
12:10 - 13:40

C
CFECMStatistics2024
13:40 - 15:20

D - Keynote
CFECMStatistics2024
15:30 - 16:20

Coffee break
16:20 - 16:50

E
CFECMStatistics2024
16:50 - 18:55

Welcome reception
19:00 - 20:30

2024-12-15

F
CFECMStatistics2024
08:45 - 10:25

Coffee break
10:25 - 10:55

G
CFECMStatistics2024
10:55 - 12:10

Lunch break
12:10 - 13:40

H
CFECMStatistics2024
13:40 - 15:20

Coffee break
15:20 - 15:50

I
CFECMStatistics2024
15:50 - 17:30

J
CFECMStatistics2024
17:40 - 18:55

Christmas Conference Dinner
20:00 - 23:00

2024-12-16

M
CFECMStatistics2024
09:10 - 10:50

Coffee break
10:50 - 11:20

N - Keynote
CFECMStatistics2024
11:20 - 12:10

Lunch break
12:10 - 13:40

O - Keynote
CFECMStatistics2024
13:40 - 14:30

P
CFECMStatistics2024
14:40 - 16:20

Coffee break
16:20 - 16:50

Q
CFECMStatistics2024
16:50 - 18:30

Closing networking drink
18:40 - 19:40

TUTORIALS, MEETINGS AND CONFERENCE DETAILS (see maps)

TUTORIALS

Three independent tutorials will take place from the 11th to the 13th of December 2024, organized within the framework of the COST Action HiTEc. The first tutorial, “Regularization methods in statistics with an application to brain imaging studies”, will be coordinated by Prof. Jaroslaw Harezlak from Indiana University, USA. The second tutorial, “Bayesian nonparametrics methods”, will be coordinated by Prof. Michele Guindani, UCLA, USA. The third tutorial, “Robust modelling of volatility and other non-negative variables”, will be coordinated by Prof. Genaro Sucarrat, BI Norwegian Business School, Norway. Further details are available on the website. Only participants who have subscribed to the tutorials can attend, either in person or virtually through the conference website.

SPECIAL MEETINGS

The *Econometrics and Statistics (EcoSta) Editorial Board* and the *CSDA and Annals of Statistical Data Science Editorial Board* meetings will take place on Friday, 13th of December 2024. Details to attend will be sent to the associate editors attending the conference in due course.

CONFERENCE DETAILS

Access

- Attendees can choose to participate virtually or in person based on their selected registration option.
- The in-person venue is King’s College London, Strand campus (Strand, London WC2R 2LS, United Kingdom).
- Instructions to access the virtual part of the conference can be found on the webpage.
- Registration will be open on Friday afternoon, from 14:00 to 18:00, during the weekend from 7:45 to 18:00 and on Monday from 8:15 to 16:30 in the Arcade of the Bush House - Central Block (ground floor).

Scientific programme and social events

- The conference will be live-streamed, with no recording available. Virtual oral presentations and posters will take place through Zoom.
- **Scientific programme:** Sessions are accessible online from the interactive schedule. The conference programme time is set in GMT. Indications to access the in-person and virtual rooms can be found on the website. The in-person participants can use S-2.23 as quiet room and to participate in virtual sessions with their laptops and headphones.
- **Coffee breaks:** The coffee breaks will last 40 minutes each (beginning 10 minutes before the times indicated in the program). These will take place at the Great Hall of the King’s Building (ground floor) and the Arcade of the Bush House - Central Block (ground floor). Participants must bring their conference badge to attend.
- **Welcome reception:** The welcome reception for registered participants will take place at the King’s building, Great Hall (Ground floor) and the Chapters (Level 2), on Saturday the 14th of December 2024 from 19:00 to 20:30 (GMT). Participants must bring their conference badge to attend the reception. Information about the welcome reception booking is embedded in the QR code on the conference badge.
- **Christmas Conference Dinner:** The Christmas Conference Dinner will take place on Sunday the 15th of December 2024, at 20:00 at the Ambassadors Bloomsbury Hotel (12 Upper Woburn Pl, London WC1H 0HX). The conference dinner is optional, and registration is required. Participants must bring their conference badge to attend the conference dinner. Information about the purchased conference dinner ticket is embedded in the QR code on the conference badge.
- **Closing networking drink:** A closing networking drink will take place on Monday, the 16th of December 2024, at 18:40 at the pub “The Last Judgment” (95 Chancery Lane, London WC2A 1DT). Participants must bring their conference badge to attend and get their voucher for a drink at the entrance.

Presentation instructions

Virtual presentations will take place through Zoom. Speakers should have a stable internet connection, and ensure their video and audio function correctly. They will share their slides when the Chair requests, present their talk, and answer the questions after the presentation. In-person speakers must copy their presentations onto the conference room laptops and share them on Zoom. Laptops are equipped with a webcam and an omnidirectional desk microphone that captures sound around the desk. Detailed instructions for speakers are available on the website. As a general rule, each speaker is assigned 20 minutes for the talk and 3-4 minutes for discussion. Strict timing must be observed.

Posters

Posters will be displayed on Zoom. In-person participants can optionally join the poster session from the room designated in the programme with their devices. Presenters should enter the poster session 10 minutes before the session starts by following the instructions for accessing it, select the breakout room with their poster code (e.g. E0123), share their poster and remain with their camera, microphone, and audio on throughout the entire session. They should also keep their chat visible.

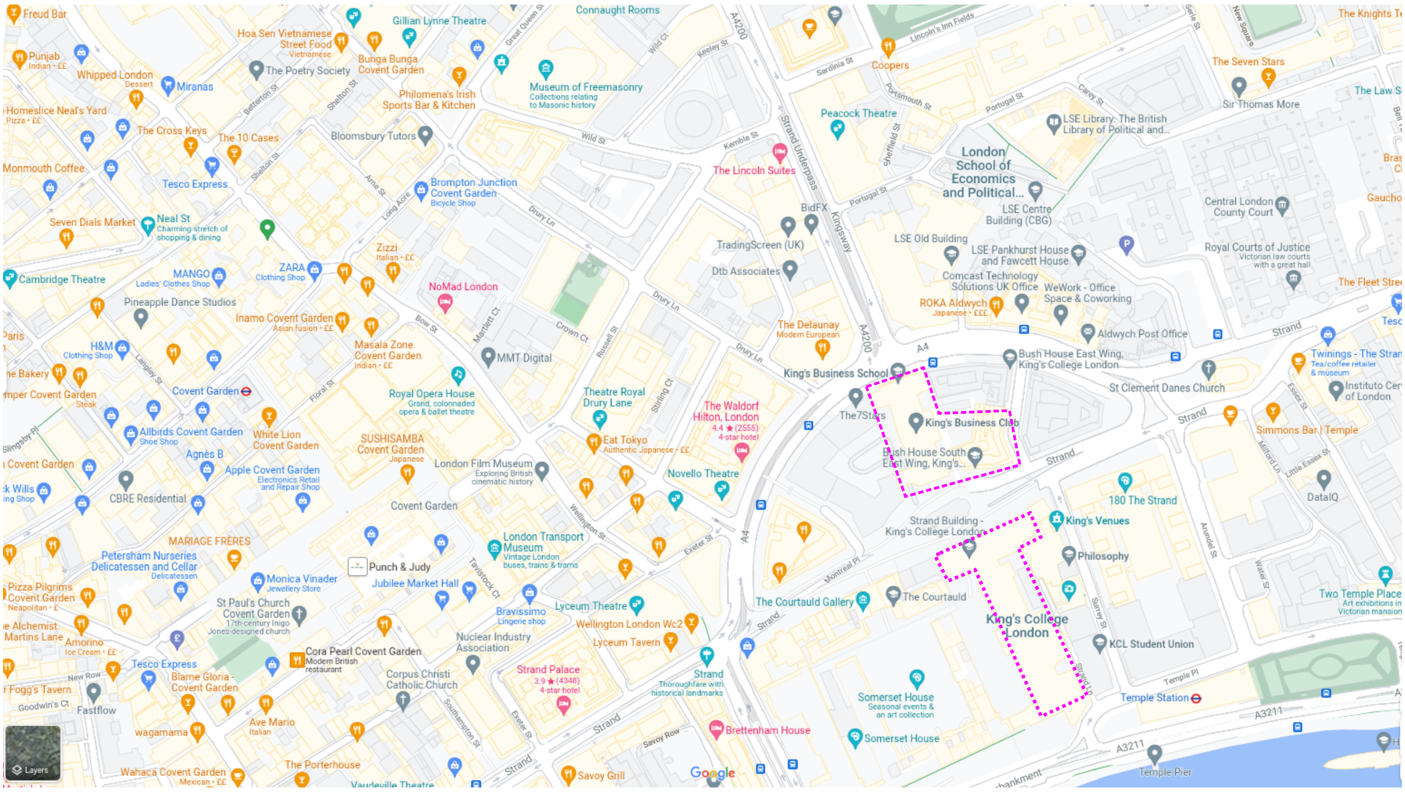
Session chairs

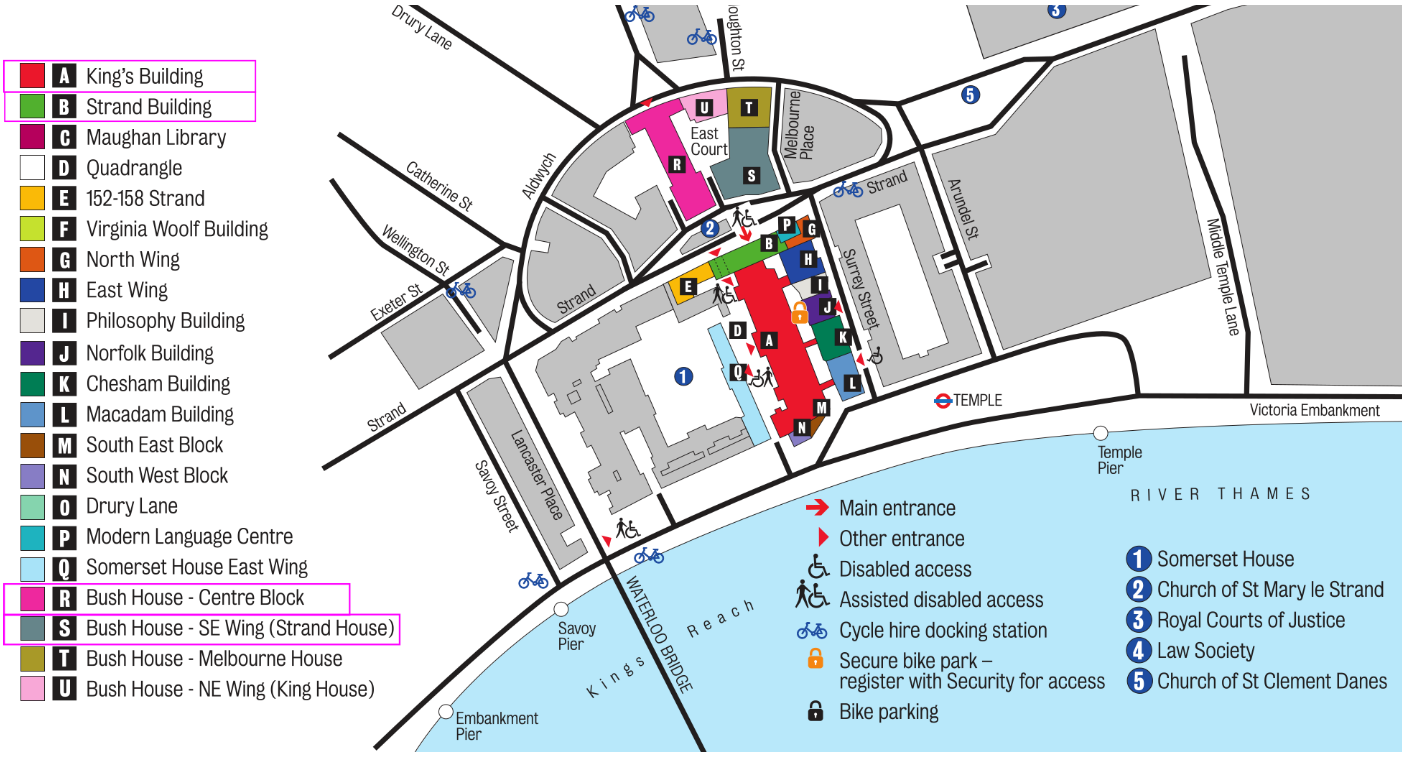
Session chairs will be responsible for introducing the session and speakers, and coordinating discussion time. A conference staff member, identified on Zoom as “Angel”, will assist online. In case of a missing or technical problem with a speaker, the Chair can move to the next speaker and return later if possible. Detailed instructions for session chairs in both virtual and hybrid sessions can be found on the website.

Test session

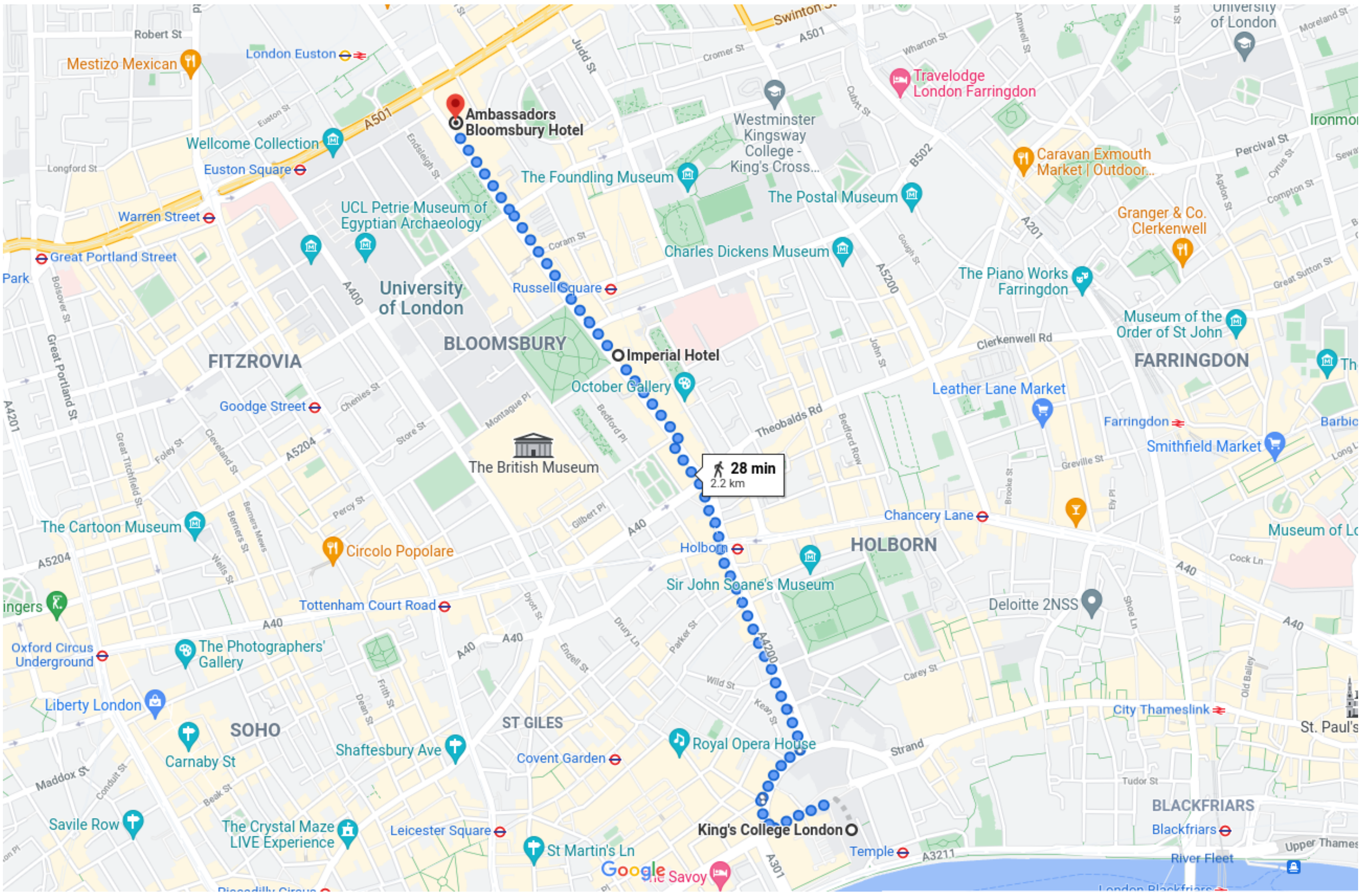
A test session is scheduled for Saturday the 7th of December 2024 from 14:00 to 14:30 (GMT). Participants will be able to virtually enter the Auditorium from the interactive programme to test presentations, video, micro and audio (e.g., through Slot A). Detailed instructions for test sessions are available on the website.

Maps of the venue and nearby area

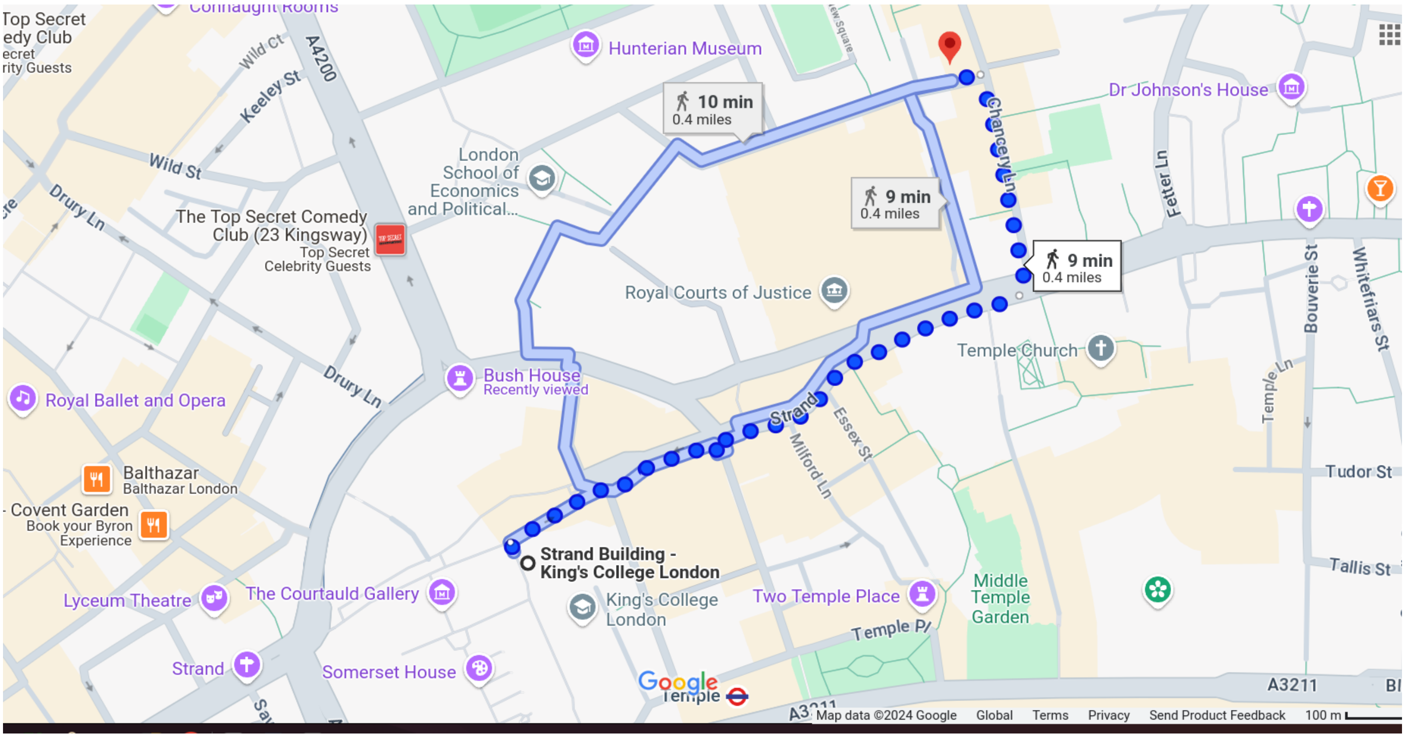




Map for the Christmas conference dinner

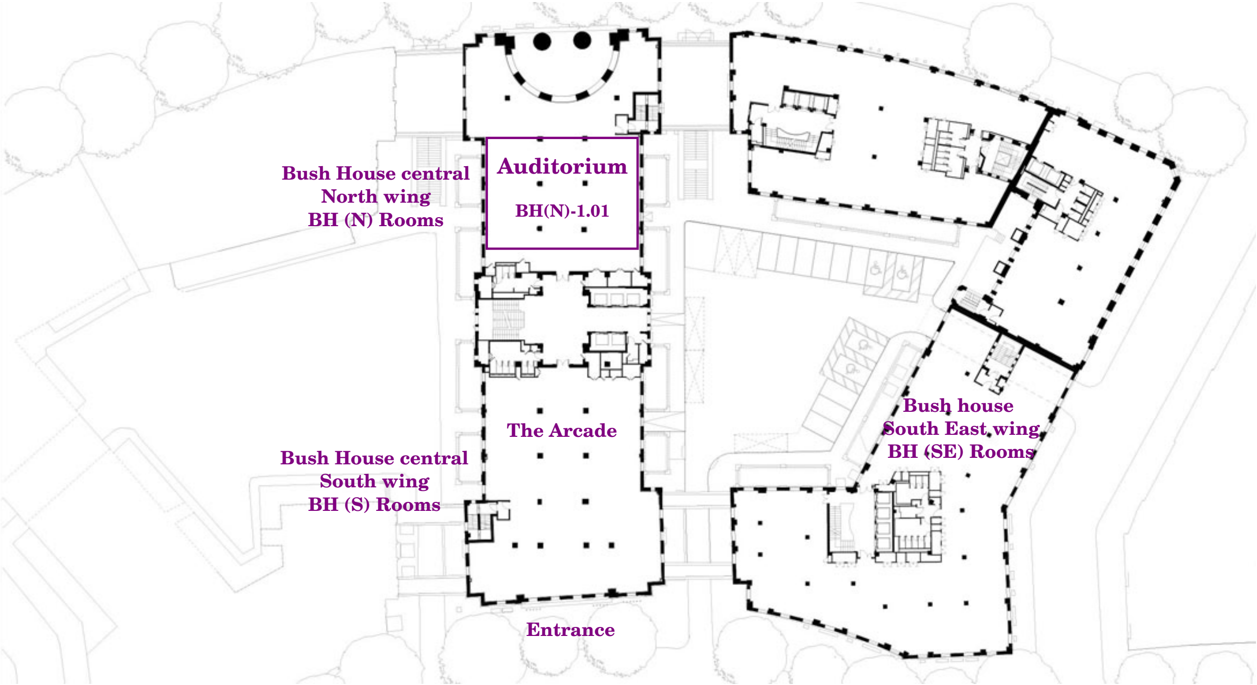


Map for the Closing drink

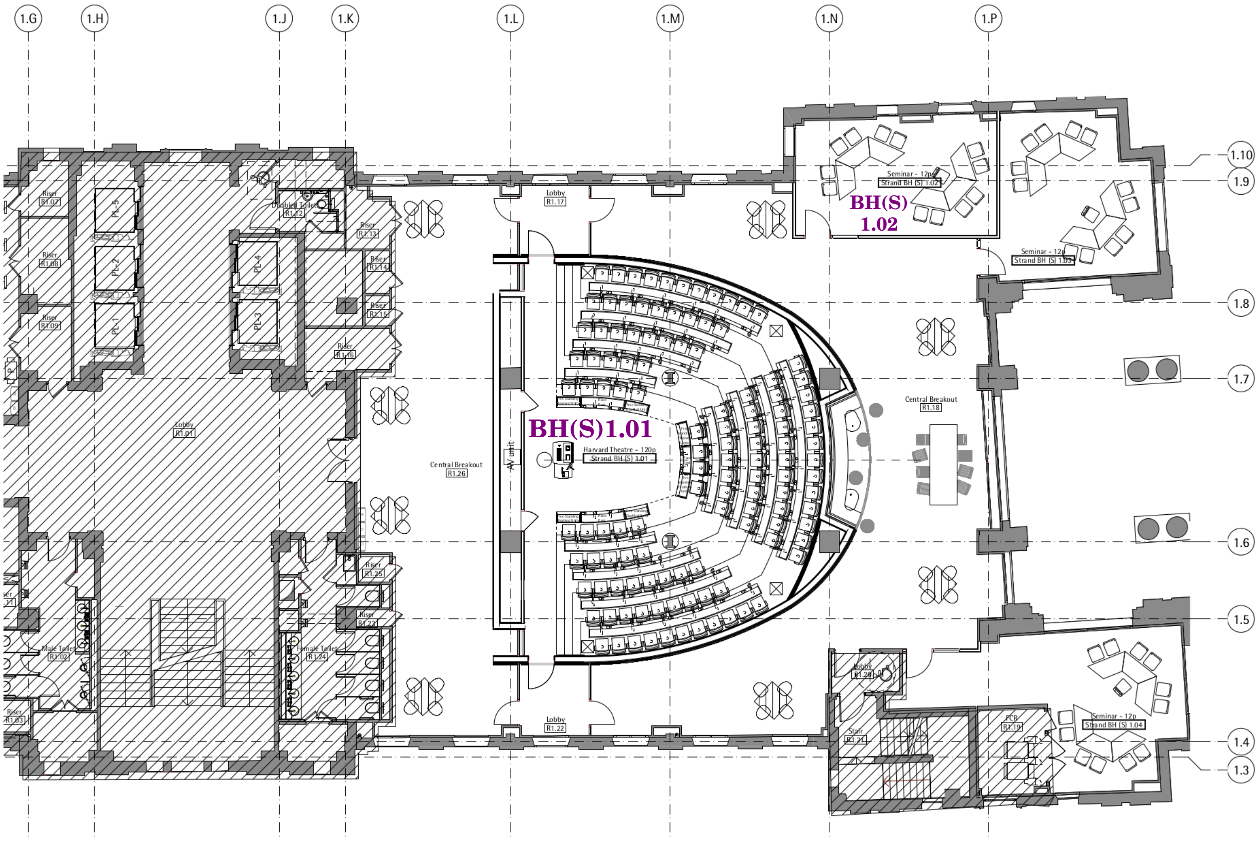


Floor maps

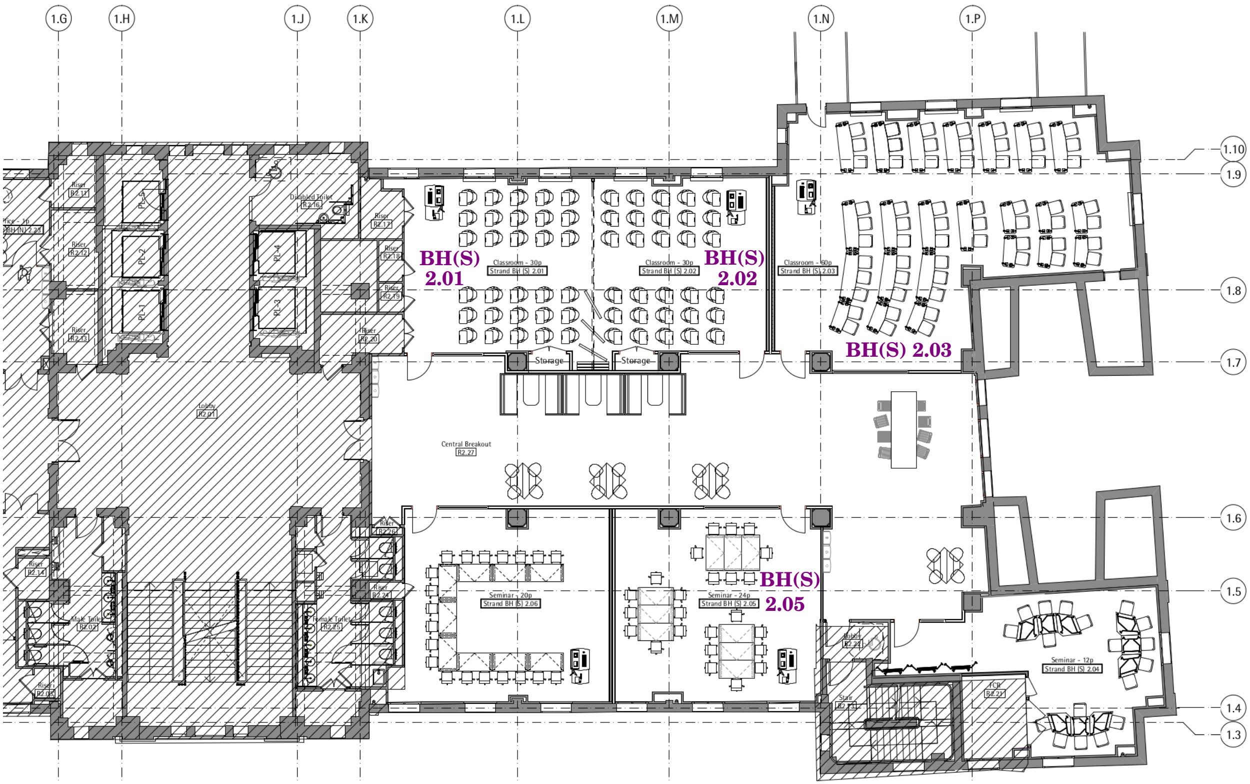
Bush House - Ground Floor



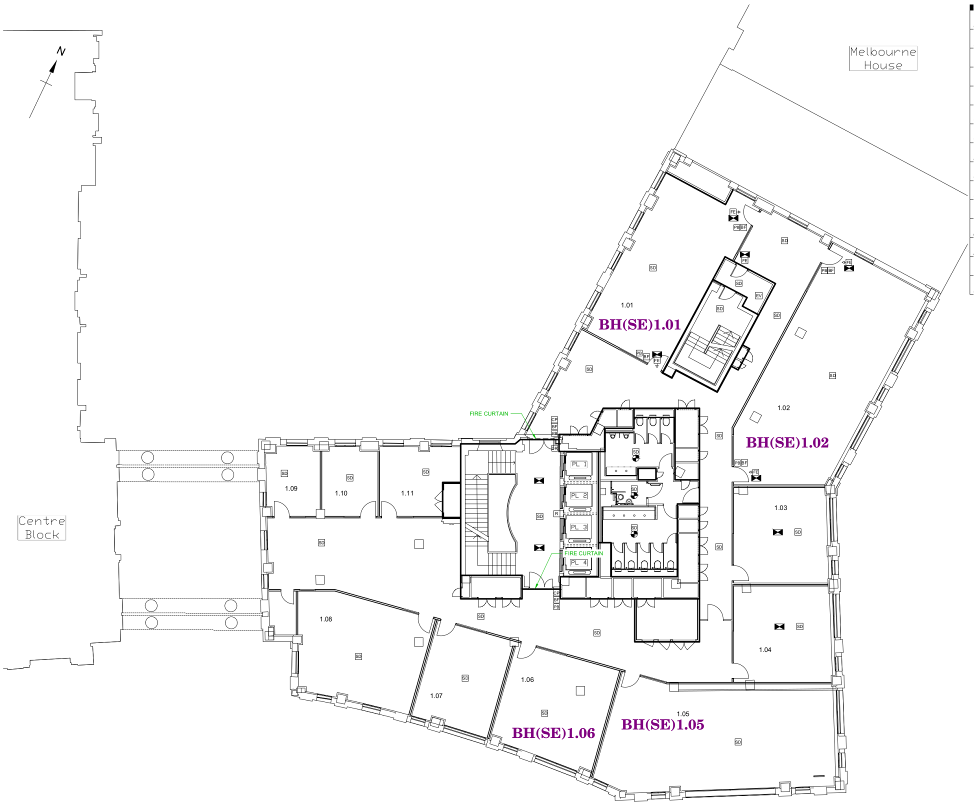
Bush House South Wing - First Floor



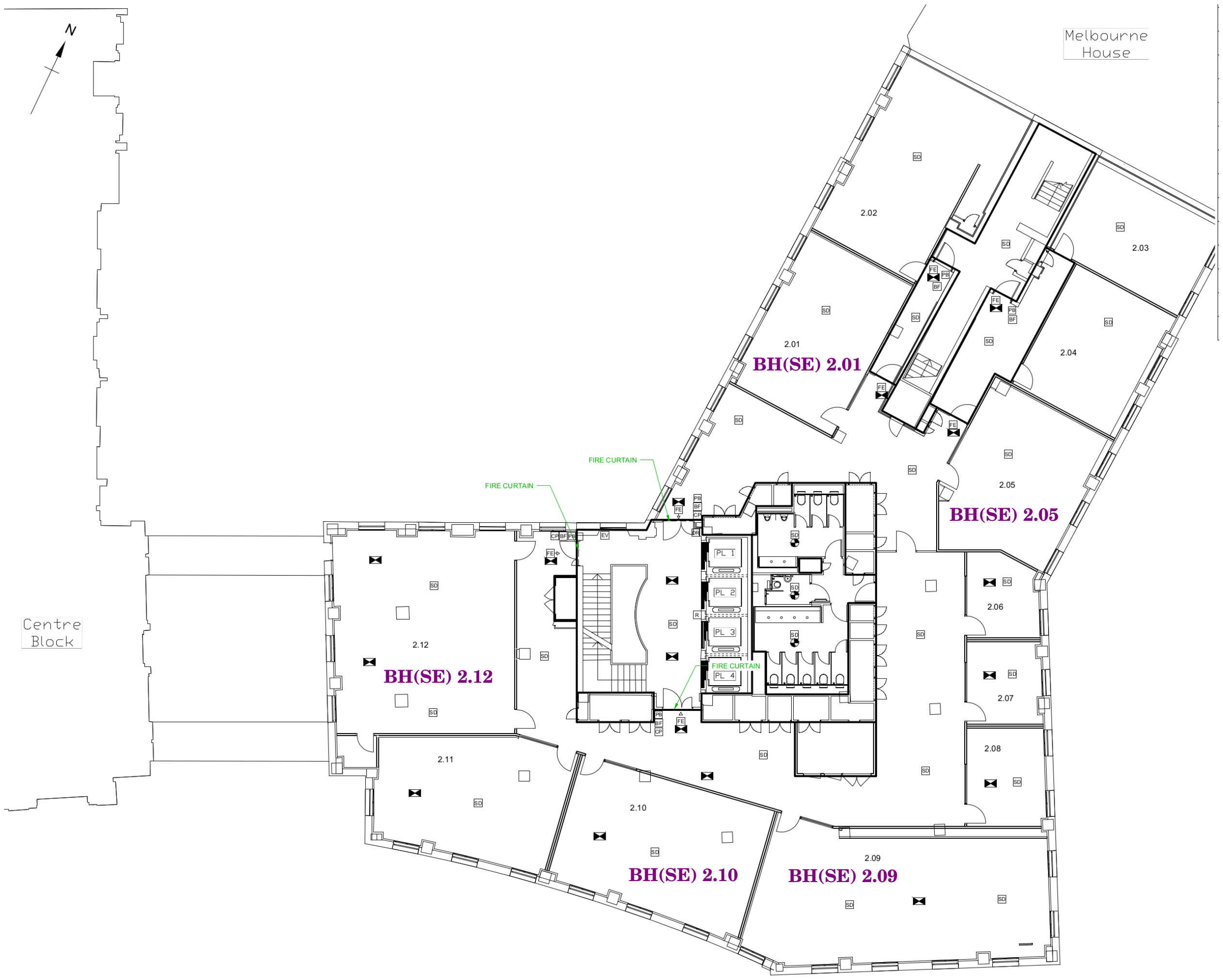
Bush House South Wing - Second Floor



Bush House South East Wing - First Floor

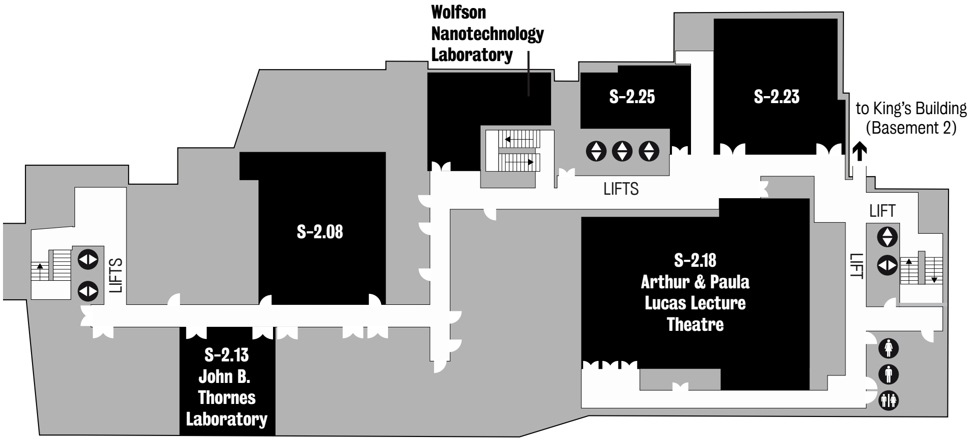


Bush House South East Wing - Second Floor



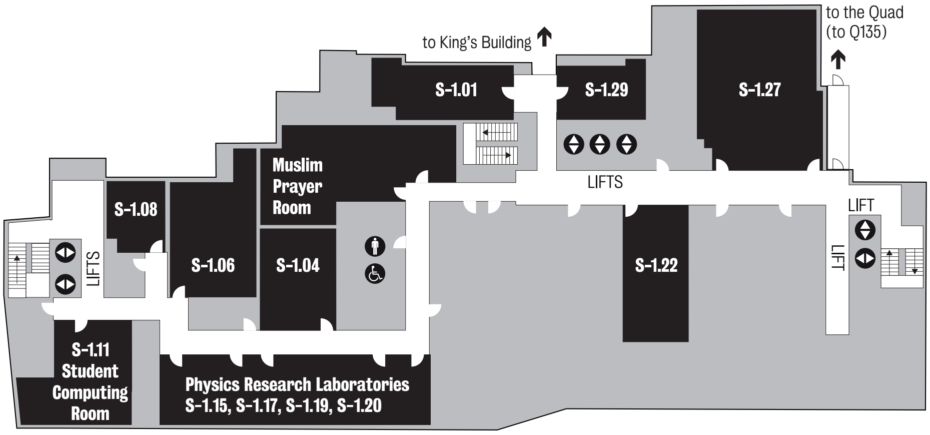
Strand Campus

Strand Building – Basement 2



Strand Campus

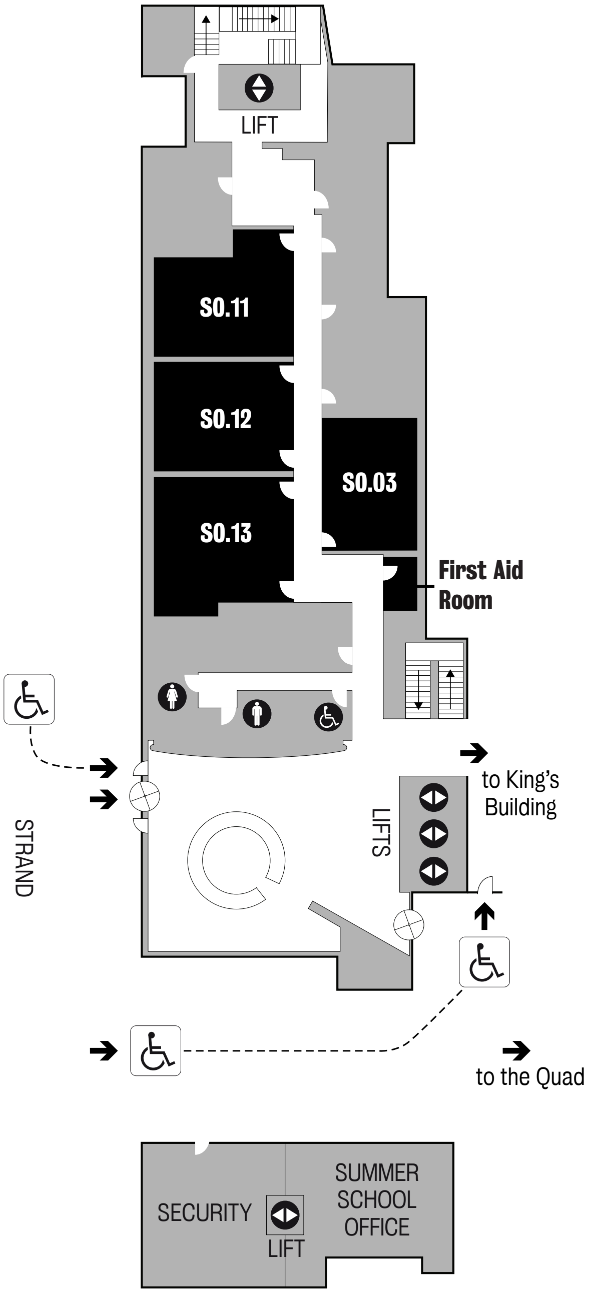
Strand Building – Basement 1



Strand Campus

KING'S College LONDON

Strand Building – Ground floor

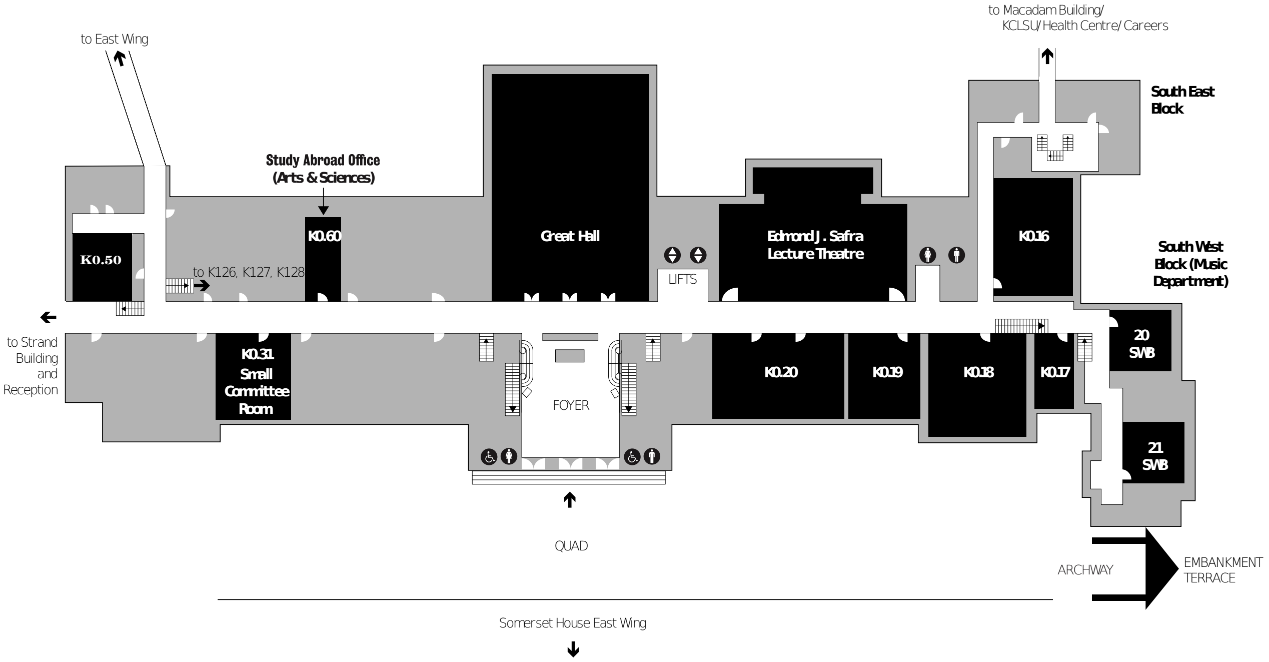


♿

The non-stepped accessible route in to the building is to the rear of the main reception area. This route is via the black gated entrance and turn left.
There is also a button-controlled self-opening door at the front of the main reception but this requires reception staff to activate it.

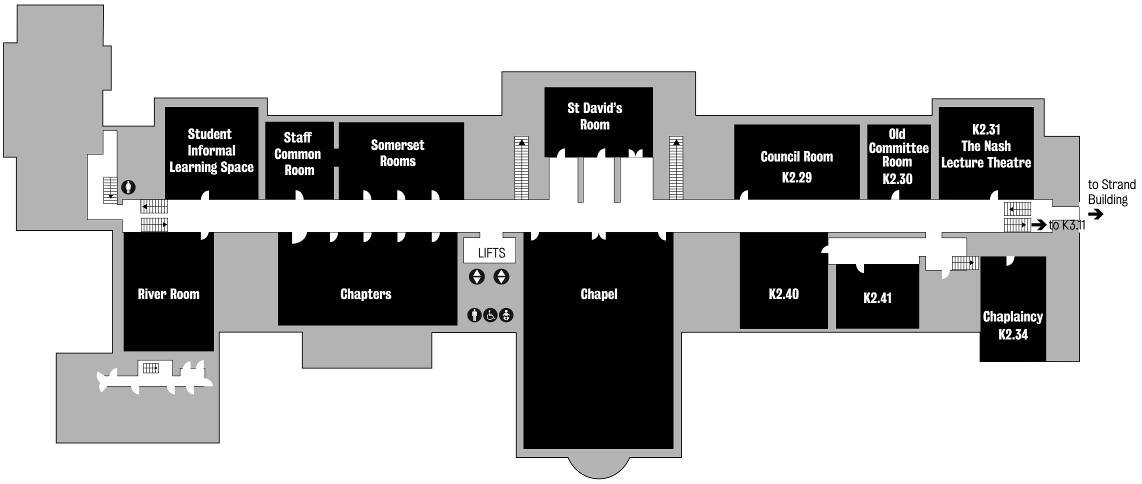
Strand Campus

King's Building – Level 0


to East Wing
to Macadam Building/ KCLSU/ Health Centre/ Careers
South East Block
Study Abroad Office (Arts & Sciences)
K0.50
K0.60
to K126, K127, K128
Great Hall
LIFTS
Edmond J. Safra Lecture Theatre
K0.16
South West Block (Music Department)
to Strand Building and Reception
K0.31 Small Committee Room
FOYER
K0.20
K0.19
K0.18
K0.17
20 SWB
21 SWB
QUAD
ARCHWAY
EMBANKMENT TERRACE
Somerset House East Wing

Strand Campus

King's Building – Level 2



PUBLICATION OUTLETS

Econometrics and Statistics (EcoSta)

http://www.elsevier.com/locate/ecosta

Econometrics and Statistics (EcoSta), published by Elsevier, is the official journal of the networks Computational and Financial Econometrics and Computational and Methodological Statistics. It publishes research papers on all aspects of econometrics and statistics and comprises two sections:
Part A: Econometrics. Emphasis is given to methodological and theoretical papers containing substantial econometrics derivations or showing potential for a significant impact in the broad area of econometrics. Topics of interest include the estimation of econometric models and associated inference, model selection, panel data, measurement error, Bayesian methods, and time series analyses. Simulations are considered when they involve an original methodology. Innovative papers in financial econometrics and its applications are considered. The covered topics include portfolio allocation, option pricing, quantitative risk management, systemic risk and market microstructure. Interest is focused as well on well-founded applied econometric studies that demonstrate the practicality of new procedures and models. Such studies should involve the rigorous application of statistical techniques, including estimation, inference and forecasting. Topics include volatility and risk, credit risk, pricing models, portfolio management, and emerging markets. Innovative contributions in empirical finance and financial data analysis that use advanced statistical methods are encouraged. The results of the submissions should be replicable. Applications consisting only of routine calculations are not of interest to the journal.
Part B: Statistics. Papers providing important original contributions to methodological statistics inspired in applications are considered for this section. Papers dealing, directly or indirectly, with computational and technical elements are particularly encouraged. These cover developments concerning issues of high-dimensionality, re-sampling, dependence, robustness, filtering, and, in general, the interaction of mathematical methods, numerical implementations and the extra burden of analysing large and/or complex datasets with such methods in different areas such as medicine, epidemiology, biology, psychology, climatology and communication. Innovative algorithmic developments are also of interest, as are the computer programs and the computational environments that implement them as a complement.
The journal consists, preponderantly, of original research. Occasionally, reviews and short papers from experts are published, which may be accompanied by discussions. Special issues and sections within important areas of research are occasionally published. The journal publishes as a supplement the Annals of Computational and Financial Econometrics.

Call For Papers Econometrics and Statistics (EcoSta)

http://www.elsevier.com/locate/ecosta

Papers presented at the conference and containing novel components in econometrics or statistics are encouraged to be submitted for publication in special peer-reviewed or regular issues of the Elsevier journal Econometrics and Statistics (EcoSta) and its supplement Annals of Computational and Financial Econometrics. Papers should be submitted using the EM Submission tool. In the EM please select as type of article the CFE conference, CMStatistics Conference or Annals of Computational and Financial Econometrics. Any questions may be directed via email to editor@econometricsandstatistics.org

Call For Papers CSDA Annals of Statistical Data Science (SDS)

http://www.elsevier.com/locate/csda

We are inviting submissions for the 1st issue of the CSDA Annals of Statistical Data Science. The Annals of Statistical Data Science is published as a supplement to the journal of Computational Statistics & Data Analysis. It will serve as an outlet for distinguished research papers using advanced computational and/or statistical methods for tackling challenging data analytic problems. The Annals will become a valuable resource for well-founded theoretical and applied data-driven research. Authors submitting a paper to CSDA may request that it be considered for inclusion in the Annals. Each issue will be assigned to several Guest Associate Editors who will be responsible, together with the CSDA Co-Editors, for the selection of papers.

Submissions for the Annals should contain a significant computational or statistical methodological component for data analytics. In particular, the Annals welcomes contributions at the interface of computing and statistics, addressing problems involving large and/or complex data. Emphasis will be given to comprehensive and reproducible research, including data-driven methodology, algorithms and software. There is no deadline for submissions. Papers can be submitted at any time. When they have been received, they will enter the editorial system immediately. All submissions must contain original unpublished work not being considered for publication elsewhere. Please submit your paper electronically using the Elsevier Editorial System: http://ees.elsevier.com/csda (Choose Article Type: Research paper, and then Select “Section IV. Annals of Statistical Data Science”).

Contents

Saturday 14.12.2024 08:45 - 09:35 Room: Auditorium Chair: Mario Peruggia Keynote talk I

Bayesian modeling in neuroimaging: Brain networks dynamics

Speaker: **Michele Guindani, University of California Los Angeles, United States**

The critical role that statistical approaches play in analyzing brain imaging data will be first highlighted, particularly for functional magnetic resonance imaging (fMRI) data. Appropriate statistical methods are necessary to handle the complexity of spatial and temporal correlations typical of brain data. More specifically, we will discuss approaches to studying dynamic brain connectivity, which seeks to understand the changing interactions between different brain regions over time. We will present two novel Bayesian approaches to capture these dynamic relationships within multivariate time series data. First, we will present a scalable Bayesian time-varying tensor vector autoregressive (TV-VAR) model, aimed at efficiently capturing evolving connectivity patterns. This model leverages a tensor decomposition of the VAR coefficient matrices at different lags and sparsity-inducing priors to capture dynamic connectivity patterns. Next, we will introduce a Bayesian framework for sparse Gaussian graphical modeling, which employs discrete autoregressive switching processes. This method improves the estimation of dynamic connectivity by modeling state-specific precision matrices, using novel prior structures to account for temporal and spatial dependencies. Throughout the talk, we will illustrate the performance of these Bayesian methods with examples from simulation studies and real-world fMRI data.

Saturday 14.12.2024 15:30 - 16:20 Room: Auditorium Chair: Matthieu Marbac Keynote talk II

Towards interpretable and trustworthy network-assisted prediction

Speaker: **Liza Levina, University of Michigan, United States** Robert Lunde, Tiffany Tang, Ji Zhu

Machine learning algorithms usually assume that training samples are independent. A network connecting the training samples tends to create dependency, reducing effective sample size but also creating an opportunity to leverage information from neighbors. Multiple prediction methods taking advantage of this opportunity have been developed, augmenting the usual node features with network features and/or neighborhood summaries. However, interpretability and inference are rarely available. Two contributions aiming to bridge this gap are covered. One is a conformal prediction method for network-assisted regression using estimated latent node positions in the network as additional features. We show that the usual conformal prediction offers finite-sample valid prediction intervals in this setting under a joint exchangeability condition and a mild regularity condition on the network statistics. The second contribution is a family of flexible network-assisted models built upon a generalization of random forests (RF+), which both achieve highly-competitive prediction accuracy and can be interpreted through importance measures, both for the features and the network. These tools help broaden the scope and applicability of network-assisted prediction for high-impact problems where interpretability and trustworthiness are essential.

Monday 16.12.2024 11:20 - 12:10 Room: Auditorium Chair: Michael Pitt Keynote talk III

Regression modelling under general heterogeneity

Speaker: **George Kapetanios, Kings College London, United Kingdom** Liudas Giraitis, Yufei Li

The aim is to introduce and analyse a setting with general heterogeneity in regression modelling. It is shown that regression models with fixed or time-varying parameters can be estimated by OLS or time-varying OLS methods, respectively, for a very wide class of regressors and noises not covered by existing modelling theory. The new setting allows the development of asymptotic theory and the estimation of standard errors. The proposed robust confidence interval estimators permit a high degree of heterogeneity in regressors and noise. The estimates of robust standard errors coincide with a well-known estimator of heteroskedasticity-consistent standard errors but are applicable to more general circumstances than just the presence of heteroscedastic noise. They are easy to compute and perform well in Monte Carlo simulations. Their robustness, generality and ease of use make them ideal for applied work. A brief empirical illustration is included.

Monday 16.12.2024 13:40 - 14:30 Room: Auditorium Chair: Kalliopi Mylona Keynote talk IV

Statistics for complex data objects - of brain structures, cell shapes and income share distributions

Speaker: **Sonja Greven, Humboldt University of Berlin, Germany**

Recent years have seen an increase in complex structured data objects that cannot be well represented by simple vectors. For such object data, the unit of observation naturally is the whole object - potentially sparsely observed and with error - examples being curve-valued, i.e., functional data, shapes, images, covariance matrices, compositions and probability densities. Functional and object data analysis aims to provide statistical methods for their analysis. Recent work will be presented that aims to transfer the flexibility and interpretability of statistical generalized additive modeling to such more complex data objects. Key ideas in our approaches are the definitions of linear or additive predictors in suitable linear spaces as well as of suitable response functions mapping to the spaces (Hilbert spaces, Riemannian manifolds or metric spaces) in which these data objects naturally live. A focus will be on running examples ranging from medicine to gender economics to illustrate all approaches.

Saturday 14.12.2024 10:05 - 12:10 Parallel Session B – CFECMStatistics2024

CI051 Room Auditorium HIGH-DIMENSIONAL TIME SERIES Chair: Tommaso Proietti

C0463: **Sparse principal component analysis for high-dimensional stationary time series**
Presenter: **Yan Liu**, Waseda University, Japan
Co-authors: Kou Fujimori, Yuichi Goto, Masanobu Taniguchi

Sparse principal component analysis (PCA) is explored for high-dimensional stationary processes. Traditional PCA is ineffective when the dimension of the time series is high. Oracle inequalities are presented for penalized PCA estimators covering a broad range of stochastic processes, including those with heavy tails. The convergence rates of these estimators are established, along with theoretical guidelines for selecting the tuning parameter. The performance of sparse PCA is illustrated through numerical simulations. Furthermore, the practical utility of sparse PCA is demonstrated using average temperature data.

C1287: **Frequency-domain estimation of dynamic factor models**
Presenter: **Giovanni Motta**, Columbia University, United States
Co-authors: Michael Eichler

The generalized dynamic factor model has become very popular in the theory and practice of large panels of time series data. The asymptotic properties of the corresponding estimators have been studied previously. Those estimators rely on Brillinger's dynamic principal components and thus involve two-sided filters, which leads to rather poor forecasting performances. A more recent study derives the asymptotic properties of a semi-parametric estimator of loadings and common shocks based on one-sided filters. However, compared to the model in the previous study, the latter model relies on the additional assumptions that the common components have rational spectral density and admit a finite autoregressive representation. Moreover, the estimator involves several time- and frequency-domain estimation steps. We propose a novel approach to estimate the common components and the common shocks directly in the frequency domain. Our approach does not rely on the assumption of rational spectral density, and our estimation method is computationally simpler and faster.

C1330: **Accurate and fast anomaly detection in industrial processes**
Presenter: **Simone Tonini**, Sant Anna School of Advanced Studies - Pisa, Italy
Co-authors: Andrea Vandin, Francesca Chiaromonte, Daniele Licari, Fernando Barsacchi

The purpose is to present a novel, simple and widely applicable semi-supervised procedure for anomaly detection in data from industrial processes, SAnD (Simple Anomaly Detection). SAnD comprises 5 steps, each leveraging well-known statistical tools, namely; smoothing filters, variance inflation factors, the Mahalanobis distance, threshold selection algorithms and feature importance techniques. To knowledge, SAnD is the first procedure that integrates these tools to identify anomalies and help decipher their putative causes. How each step contributes to tackling technical challenges that practitioners face when detecting anomalies is shown in industrial contexts, where signals can be highly multicollinear, have unknown distributions, and intertwine short-lived noise with the long(er)-lived actual anomalies. The development of SAnD was motivated by a concrete case study from the industrial partner, which is used to show its effectiveness. The performance of SAnD is also evaluated by comparing it with a selection of semi-supervised methods on public datasets from the literature on anomaly detection. SAnD is concluded to be effective, broadly applicable, and outperforms existing approaches in both anomaly detection and runtime.

CO170 Room S-1.01 HITEC: ADVANCES IN FINANCIAL ECONOMETRICS Chair: Genaro Sucarrat

C0878: **On the credibility of the 2015 Paris agreement and effectiveness of climate policies**
Presenter: **Susana Campos Martins**, University of Oxford, United Kingdom

The aim is to propose a novel approach based on financial asset returns to empirically investigate how the response of global financial markets to climate change concerns has changed over time. Leveraging a data-driven approach to select time indicators, it is found that most climate-driven global common volatility of the oil and gas industry is material only in the period post-Paris Agreement. Policies designed to reduce greenhouse gas emissions and mitigate the effects of climate change by ramping up investment in renewable energy and implementing carbon taxes and government programs that focus on the energy transition all seem to be creating turmoil in the oil and gas industry. This is contrary to green markets, where such climate policies make green asset returns less volatile, so the investments there are less risky. The contributions shed light on the credibility of international cooperation under the Paris Agreement and the effectiveness of climate policies under nationally determined contributions. By being more ambitious, transparent, and accountable, we find that the Paris Agreement appears to be sending the right signals to financial markets concerning the energy transition since its adoption. Ever since, climate concerns seem to be disrupting the oil and gas industry at an increasing rate, while green assets tend to be more stable, safer investments.

C0981: **An economic evaluation of exchange rates higher order moments timing**
Presenter: **Francesco Violante**, IESEG School of Management, France
Co-authors: Stefano Grassi

The short-term predictive ability of empirical exchange rate models is evaluated within a framework that allows for volatility, correlation, (co-)skewness and (co-)kurtosis timing. The Mixture of (Student's) t by Importance Sampling Weighted Expectation-Maximization (MitISEM) is adapted to the estimation of multivariate Skew-t models featuring observation-driven time-varying covariance matrix, as well as asymmetry parameter and degrees of freedom. Compared to model-specific Markov chain Monte Carlo (MCMC) procedures, the MitISEM is more general, parallelizable, and it is considerably faster. Additionally, it provides the model's marginal likelihood, which is useful for model selection. Using GBP, EUR and JPY exchange rates relative to the USD, the economic value of these models' forecasting power is assessed by evaluating the impact of predictable changes in the conditional moments of foreign exchange returns on the performance of dynamic allocation strategies.

C0856: **Testing for breaks in the conditional mean based on the estimating function approach**
Presenter: **Jean-Michel Zakoian**, CREST, France
Co-authors: Christian Francq, Lorenzo Trapani

The estimating function approach is particularly attractive for time series models where the dynamics are not fully specified, but the conditional mean is assumed to be a given parametric function of past observations. In many financial applications, however, the conditional mean may undergo a structural change. A class of cumulative sum, CUSUM, statistics is proposed to detect breaks in the conditional mean under weak assumptions. This procedure depends on the choice of a sequence of weights, leading to a potentially infinite number of consistent tests, and it is shown that the best test is related to Godambe's optimal estimator, also discussing data-driven procedures for this optimal choice of weights. Inference is studied in the presence of a changepoint, and the case is also studied where the conditional mean is misspecified, developing heteroskedasticity and autocorrelation consistent (HAC) versions of the test. Results are illustrated using Monte Carlo experiments and real financial data.

C0521: **Volatility prediction under misspecification**
Presenter: **Genaro Sucarrat**, BI Norwegian Business School, Norway

The volatility models used in practice are unlikely to equal the data-generating process (DGP). Accordingly, models that are valid under misspecification are of great importance. Exact, general and mild conditions are established under which a large class of volatility prediction specifications exists. Crucially, the specifications within the class generate volatility predictions that are weakly identified for volatility under misspecification. Next, a consistent and asymptotically normal estimator that is valid under dependence of unknown form is derived. The volatility prediction specifications considered in more detail are modifications of the log-ARCH-X model. The specifications are highly interpretable and versatile and accommodate zero returns (in contrast to the classic log-ARCH specification), short-term and long-term persistence, asymmetry, volatility proxies and additional covariates. Since the volatility specifications are in logs, the inference is standard under the nullity of the parameters, and the positivity of the volatility predictions is guaranteed. In the simulation experiments, the predictions are both unbiased and identified for the benchmark model, whereas in the empirical illustration, the volatility predictions compare well with those of the benchmark volatility model.

C1475: **A HAR-based stochastic volatility model for leverage propagation**
Presenter: **Helena Veiga**, Universidad Carlos III de Madrid, Spain
Co-authors: J Miguel Marin, Eva Romero

The aim is to propose a stochastic volatility model that uses a heterogeneous autoregressive process to capture the persistence of leverage over time. The properties of the model are analyzed by simulation in terms of leverage and leverage propagation using a recent concept in the field, and it is found that the model can generate both effects. Data cloning is also introduced for parameter estimation, which provides accuracy and computational efficiency in finite samples. Empirical analysis shows the proposal has good in-sample and out-of-sample performances across different financial return series. This makes it an effective and simple tool for capturing leverage dynamics in financial markets.

CO005 Room S-1.04 STOCHASTIC PROCESSES: THEORY AND APPLICATIONS Chair: Lorenzo Mercuri

C0482: **Profile quasi-likelihood inference for SDE with mixed effects**
Presenter: **Hiroki Masuda**, University of Tokyo, Japan
Co-authors: Maud Delattre

Mixed-effects models play a pivotal role across various scientific domains, facilitating the precise analysis of repeated observations among individuals. Recent advancements propose random-effect modeling based on the generalized hyperbolic distribution to better accommodate variability. Utilizing normal variance-mean mixture-type random effects is considered in a class of stochastic differential equations (SDE) with mixed effects. The statistical framework allows for the exploration of a wider class of mixed-effects diffusion models compared to previous literature. A novel parameter estimation method is proposed, and theoretical insights are provided into the asymptotic behavior of the estimators. The estimation method diverges from the quasi-likelihood approach, offering a more accessible numerical procedure while sacrificing some efficiency compared to maximum likelihood estimation. This trade-off ensures stability and ease of implementation in high-frequency frameworks.

C0655: **Efficient drift parameter estimation for ergodic solutions of backward SDEs**
Presenter: **Teppei Ogihara**, University of Tokyo, Japan
Co-authors: Mitja Stadje

Efficient drift parameter estimation is explored for ergodic solutions of backward stochastic differential equations (BSDEs). Traditional methods for estimating parameters in stochastic differential equations (SDEs) often assume known parametric forms for the diffusion coefficient. However, the diffusion coefficient is not parametrized nor observed when the process is BSDE. A maximum likelihood type estimation method is proposed that leverages discrete observations of the BSDE to estimate the drift parameter in the presence of an unknown diffusion coefficient. The approach involves constructing quasi-log-likelihood functions using discrete-time observations and employing ergodic properties of the underlying processes. Under appropriate smoothness and non-degeneracy conditions, the maximum likelihood estimators (MLEs) for the drift parameter are demonstrated to achieve asymptotic normality, ensuring reliable and consistent parameter estimation as the sample size increases. Numerical experiments are conducted to validate the theoretical results and to illustrate the practical performance of the proposed estimators.

C0755: **Yet another approximation for the total claims amount using the Weibull distribution**
Presenter: **Alessandro Barbiero**, Universita degli Studi di Milano, Italy

The accurate evaluation of the distribution of a compound sum is a crucial task in actuarial science and operational risk management. For non-life insurance companies, the total claims amount over a specific period can be represented as $S_N = X_1 + \ldots + X_N$, where N denotes the number of occurring claims and X_i the i-th claim size ($i = 1,\ldots,N$). The X_i's are assumed to be iid positive random variables, typically continuous, and N is a counting random variable independent of the X_i's. The evaluation of the distribution of S_N is challenging: only in a few situations one can derive it analytically; in the other cases, one needs to resort to numerical methods, Monte Carlo simulations, or discrete/continuous approximations. Focusing on this latter technique, one common approach is to approximate the distribution of S_N using normal, normal-power or translated Gamma distributions, whose parameter values are obtained by matching the same-order moments. An approximation of the total claims amount distribution by a three-parameter Weibull distribution is introduced, discussed, and assessed. This assessment considers different combinations of distributions for the claim frequency and size. The availability of relatively easy expressions for the first three non-central moments facilitates its use. However, care should be taken as the level of approximation might be unsatisfactory for some parts of the distribution under certain circumstances.

C1089: **The greenium term structure**
Presenter: **Edit Rroji**, Universita' degli studi di Milano-Bicocca, Italy
Co-authors: Lorenzo Mercuri, Ilaria stefani

Green bonds provide financial backing for low-carbon initiatives and facilitate the transition towards a greener economy. The greenium effect refers to the potential premium bondholders are willing to pay to invest in green securities compared to investments with similar characteristics such as maturity, coupon rate, and issuer credit profile. Despite the interest of recent literature on this topic, the determinants and dynamics of the greenium effect remain inadequately understood, particularly concerning its term structure and geographical dependencies. A mathematical framework is proposed, employing an autoregressive model where the error term is conditionally gamma-distributed. Leveraging likelihood estimation techniques and market bond prices, the methodology aims to establish a comprehensive framework for the term structure of the greenium effect.

C1329: **Parameter inference for hypo-elliptic diffusions under a weak design condition**
Presenter: **Yuga Iguchi**, University College London, United Kingdom
Co-authors: Alexandros Beskos

The problem of parameter estimation is addressed for degenerate diffusion processes defined via the solution of stochastic differential equations (SDEs) with a diffusion matrix that is not full-rank. For this class of hypo-elliptic diffusions, recent works have proposed contrast estimators that are asymptotically normal, provided that the step-size in-between observations $\Delta = \Delta_n$ and their total number n satisfying the classical high-frequency setting with $\Delta = o(n^{-1/2})$. This latter restriction requires a so-called rapidly increasing experimental design. This limitation is overcome, and a general contrast estimator is developed, satisfying asymptotic normality under the weaker design condition $\Delta = o(n^{-p/2})$ for general integer p greater than 2. Such a result has been obtained for elliptic SDEs in the literature, but its derivation in a hypo-elliptic setting is highly non-trivial. Numerical results are provided to illustrate the advantages of the developed theory.

CO350 Room K0.16 STATISTICAL ANALYSIS OF NETWORKS AND APPLICATIONS Chair: Wendy Meiring

C1604: **Networks in neuroscience: Functional and structural brain connectivity**
Presenter: **Wendy Meiring**, University of California Santa Barbara, United States

Two applications of networks are reviewed when estimating functional and structural brain connectivity. Part 1: Resting-state functional brain connectivity quantifies the similarity in neuronal activity (firing) across brain regions/voxels within brain regions over time. Each brain region consists of voxels at which dynamic signals are acquired via neuroimaging measurements, such as blood-oxygen-level-dependent (BOLD) signals in functional magnetic resonance imaging (fMRI). Pearson correlations and similar metrics are frequently adopted by neuroscientists to estimate inter-regional functional connectivity, often after averaging of signals across voxels within each region; however, careful attention needs to be paid to account for inter- and intra-regional spatiotemporal noise on these estimates. Part 2: Network concepts used in structural (anatomical) brain connectivity estimation are reviewed based on structural MRI scans.

C1270: **A convex formulation of covariate-adjusted Gaussian graphical models via natural parametrization**
Presenter: **Ruobin Liu**, University of California, Santa Barbara, United States
Co-authors: Guo Yu

Gaussian graphical models are widely used to recover the conditional independence structure among random variables. Recently, several key advances have been made to exploit an additional set of variables to estimate the graphical model of the variables of interest better. For example, in co-expression quantitative trait locus studies, both the mean expression level of genes and their pairwise conditional independence structure may be adjusted by genetic variants local to those genes. Existing methods to estimate covariate-adjusted graphical models either allow only the mean to depend on covariates or suffer from poor scaling assumptions due to the inherent non-convexity of simultaneously estimating the mean and precision matrix. A convex formulation that jointly estimates the covariate-adjusted mean and precision matrix is proposed by utilizing the natural parametrization of the multivariate Gaussian likelihood. This convexity yields theoretically better performance as the sparsity and dimension of the covariates grow large relative to the number of samples. The theoretical results are verified with numerical simulations and a reanalysis of a study of glioblastoma multiforme is performed.

C1406: **Online graph topology learning from matrix-valued time series**
Presenter: **Yiye Jiang**, University of Grenoble Alpes, France

The focus is on the statistical analysis of matrix-valued time series, where data is collected over a network of sensors, typically at spatial locations, over time. Each sensor records a vector of features per time instant. The goal is to identify the dependency structure among these sensors and represent it as a graph. When one feature per sensor is observed, vector auto-regressive (VAR) models are commonly used to infer Granger causality, forming a causal graph. The first contribution extends VAR models to matrix-variate models for graph learning. Two online procedures are proposed for low and high dimensions, enabling rapid updates of estimates as new samples arrive. In high dimensions, a novel Lasso-type is developed, and its homotopy algorithms are introduced for online learning. An adaptive tuning procedure for the regularization parameter is provided. Given that detrending is often required for applying auto-regressive models but infeasible in online settings, the proposed AR models are augmented by incorporating trend as an additional parameter, with a particular focus on periodic trends. The adapted algorithms can, therefore, simultaneously learn the graph and trend from streaming data. Numerical experiments with synthetic and real data demonstrate the effectiveness of these methods.

C1590: **Networks inference with (quasi-)analytic wavelets**
Presenter: **Sophie Achard**, CNRS LJK, France
Co-authors: Irene Gannaz

In the general setting of long-memory multivariate time series, the long-memory characteristics are defined by two components. The long-memory parameters describe the autocorrelation of each time series. The long-run covariance measures the coupling between time series and general phase parameters. It is of interest to estimate the long-memory, long-run covariance and general phase parameters of time series generated by this wide class of models, although they are not necessarily Gaussian nor stationary. This estimation is thus not directly possible using real wavelets decomposition or Fourier analysis. The purpose is to define an inference approach based on a representation using quasi-analytic wavelets. It is first shown that the covariance of the wavelet coefficients provides an adequate estimator of the covariance structure, including the phase term. Consistent estimators based on a local Whittle approximation are then proposed. Simulations highlight a satisfactory behavior of the estimation of finite samples on multivariate fractional Brownian motions. An application on a real neuroscience dataset is presented, where long-memory and brain connectivity are inferred.

C1684: **High-dimensional semiparametric skew-elliptical copula graphical models**
Presenter: **Gabriele Di Luzio**, Sapienza, University of Rome, Italy

A semiparametric approach called elliptical skew-(S)KEPTIC is proposed for efficiently and robustly estimating non-Gaussian graphical models. Relaxing the assumption of meta-elliptical distribution into the family of meta skew-elliptical distributions that accommodates a skewness component, we derive a new estimator that is an extension of the SKEPTIC estimator. This extension is based on semiparametric Gaussian copula graphical models and applies to skew-elliptical copula graphical models. Theoretically, we demonstrate that the elliptical skew-(S)KEPTIC estimator achieves robust parametric convergence rates in both graph recovery and parameter estimation. We conduct numerical simulations to verify the reliable graph recovery performance of the elliptical skew-(S)KEPTIC estimator. Finally, the new method is applied to the daily log-returns of the stocks of the S&P 500 index and shows better interpretability compared to Gaussian copula graphical models.

CO279 Room K0.19 COMPLEX ENVIRONMENTAL DATA AND MODELING (COENV) Chair: Nicola Pronello

C0308: **Scalable additive Gaussian process regression using Vecchia approximations**
Presenter: **Isa Marques**, The Ohio State University, United States
Co-authors: Paul Wiemann, Matthias Katzfuss

Generalized additive models (GAMs) have gained widespread popularity for their ability to link a set of covariates to a response through potentially non-linear effects. The recent integration of Gaussian processes (GPs) into this framework has further advanced modeling capability. In a configuration restricted to additive, one-dimensional GPs employing stationary and isotropic Matern covariance functions, it has been shown that the posterior mean and variance computations are achievable in quasilinear complexity. To overcome these restrictions while maintaining computational efficiency, a novel method is presented utilizing Vecchia approximations of the latent additive functions. This approach facilitates efficient computation of posterior mean and variance in more complex scenarios. Notably, it accommodates interactions and non-Matern covariance functions, which are pivotal in machine learning applications.

C0642: **Confounding adjustment with spatiotemporal data**
Presenter: **Carlo Zaccardi**, University of Chieti & Pescara, Italy
Co-authors: Pasquale Valentini, Luigi Ippoliti

In the context of epidemiological data with a spatiotemporal structure, the association between exposure and outcome is of interest. Besides, spatiotemporal confounders may exist, i.e. factors that influence the outcome and have spatial and temporal trends similar to those of the exposure. The problem known as spatiotemporal confounding arises when these confounders are not measured. For instance, if the association between exposure to fine particulate matter and mortality is of interest, weather variables and influenza outbreaks are examples of confounding factors, with the difference that the first ones are measured but the second ones are not. However, not accounting for these unknown factors leads to distorted conclusions about the exposure's effect; indeed, confounding is considered a major challenge in epidemiology. Also, the shape of the association is generally not known and may vary with space and time. Henceforth, to recover the association of interest, the statistical model must account for its possible non-linearity and for spatiotemporal confounding as well. To this end, a Bayesian approach is proposed where the regression coefficients are allowed to vary with space and time. The proposed method can capture potential non-linearities and alleviate the spatiotemporal confounding bias.

C0696: **Unlocking insights into UK rivers using statistics and data analytics**
Presenter: **Ruth ODonnell**, University of Glasgow, United Kingdom
Co-authors: Marian Scott, Claire Miller, Ionut Paun

In the world of continually expanding environmental data streams, it is essential to fully harness the power of these data in order to understand the current status and changing dynamics of our natural resources. The purpose is to discuss the current work on the MOT4 Rivers (NE/NE/X01620X/1) project, which aims to explore how pollution and climate change impact freshwater ecosystems. To date, much of the understanding of the impact of pollutant exposures and changing conditions on the river ecosystems has been limited by viewing individual biological and chemical drivers in isolation and not considering their combined effects. The aim is to employ cutting-edge data analysis techniques to integrate a range of both environmental and ecological measurements and thus maximize the insights on the pressures faced by UK river systems. However, while data from routine monitoring for European and UK regulatory purposes are available at the national scale, include several thousands of sites, and have been observed for more than ten years, they are still sparse and misaligned spatially and temporally. This leads to an array of statistical challenges in exploring the changing status of the ecosystems.

C1082: **Estimating graphical models varying on a spatial network for water quality assessment**
Presenter: **Rosaria Ignaccolo**, University of Turin, Italy
Co-authors: Nicola Pronello, Alex Cucco, Vito Frontuto, Natalia Golini, Luigi Ippoliti

The purpose is to illustrate how graphical models that vary across a spatial network provide an effective way to accurately depict the conditional dependence relationships among water pollutants and other features observed over a fluvial network. Once all graphs - one for each site - have been estimated, their analysis enables tracking changes in the ecological and chemical status of water bodies along the fluvial network. By considering distances among graphs, it is possible to cluster spatial sites and identify potentially critical zones. Additionally, graph analytics can be used to detect spatially varying clusters of variables throughout the fluvial network. The motivating case study focuses on assessing the quality of water bodies in the Italian region of Piedmont.

CO376 Room K0.50 FACTORIAL DESIGNS UNDER MODEL UNCERTAINTY Chair: Steven Gilmour

C1243: **An overview of the *Q_B*-optimality criterion**
Presenter: **Steven Gilmour**, KCL, United Kingdom

The *Q_B* criterion provides a link between the optimal and classical designs of experiments. It was developed from an approximation to an optimal design approach, averaged over many models, but several classical criteria, especially generalizations on minimum aberration, can be shown to be special cases of *Q_B*-optimality. Searching for a *Q_B*-optimal design has many advantages over some more recent classical-type approaches but is not yet widely used. The purpose is to show how simple and effective it is to use *Q_B* in practice, and some recently produced software for finding *Q_B*-optimal designs will be described.

C0334: **Constructing two-level QB-optimal screening designs using mixed-integer programming and heuristic algorithms**
Presenter: **Alan Vazquez**, Tecnologico de Monterrey, Mexico
Co-authors: Peter Goos, WengKee Wong

Two-level screening designs are widely applied in the manufacturing industry to identify influential factors of a system. These designs have each factor at two levels and are traditionally constructed using standard algorithms, which rely on a pre-specified linear model. Since the assumed model may depart from the truth, two-level QB-optimal designs have been developed to provide efficient parameter estimates for several potential models. These designs also have an overarching goal, which is that models that are more likely to be the best for explaining the data are estimated more efficiently than the rest. However, there is no effective algorithm for constructing them. Two methods are presented: A mixed-integer programming algorithm that guarantees convergence to the two-level QB-optimal designs and a heuristic algorithm that employs a novel formula to find good designs in short computing times. Using numerical experiments, the mixed-integer programming algorithm is shown to be attractive for finding small optimal designs, and the heuristic algorithm is the most computationally-effective approach for constructing both small and large designs when compared to benchmark heuristic algorithms.

C1173: **Optimal two-level designs under model uncertainty**
Presenter: **Pi-Wen Tsai**, National Taiwan Normal University, Taiwan
Co-authors: Steven Gilmour

Two-level designs are widely used for screening experiments, with the goal of identifying a few active factors that have major effects. The model-robust Q_B criterion is applied for the selection of optimal two-level designs without the requirement of level balance and pairwise orthogonality. A coordinate exchange algorithm is provided for the construction of Q_B-optimal designs for the first-order maximal model and second-order maximal model, and it is demonstrated that different designs are recommended based on different experimenters' prior beliefs. Additionally, the definition of Q_B-criterion is extended to regular and irregular block designs and the relationship between this new criterion and the aberration-type criteria for blocks is studied. Some trade-offs between orthogonality and confounding will lead to different choice of block designs. Some new classes of model-robust designs that respect experimenters' prior beliefs have been found.

C0913: **A complete catalog of D- and A-optimal designs with up to 20 runs**
Presenter: **Mohammed Saif Ismail Hameed**, KU Leuven, Belgium
Co-authors: Eric Schoen, Jose Nunez Ares, Peter Goos

The literature on two-level D- and A-optimal designs for the main-effects model is very exhaustive for run sizes that are multiples of four. This is due to the fact that complete catalogs of D- and A-optimal designs exist for run sizes that are multiples of four. However, for run sizes that are not multiples of four, there are no such catalogs, and experimenters resort to heuristic optimization algorithms to create designs. This approach has multiple weaknesses. First, it requires computing time. Second, heuristic optimization algorithms often fail to return a truly optimal design. Third, even in the event the design produced is truly optimal for the main-effects model, it often exhibits substantial aliasing between the main effects and the two-factor interactions as well as among the two-factor interactions. The purpose is to explain how to enumerate complete catalogs of D- and A-optimal main-effects designs for run sizes that are not multiples of four and how to select the best of these designs in terms of aliasing between the main effects and the two-factor interactions and among the two-factor interactions. As a result, the use of heuristic optimization can be avoided for most optimal design problems where the run size is at most 20 in the event the primary interest of the experimenter is in a main-effects model and statistical software can provide a minimally aliased D- and A-optimal design instantaneously.

C0262: **Optimal designs under model uncertainty**
Presenter: **Xietao Zhou**, KCL, United Kingdom
Co-authors: Steven Gilmour

Traditional optimal designs are optimal under the pre-specified model. When the final fitted model differs from the pre-specified model, traditional optimal designs may cease to be optimal, and the corresponding parameter estimators may have larger variances. The Q_B criterion has been proposed to consider hundreds of alternative models that could appear for multifactor designs in the sense that the Q_B-optimal design shall give better parameter estimation in more alternative models than the traditional optimal designs. Recently, an alternative parameterization of factorial designs called baseline parameterization has been considered in the literature. It has been argued that such a parameterization arises naturally if there is a null state of each factor, and the corresponding optimal design has been explored. The basic framework of the Q_B criterion and how it could be extended to the baseline parameterization is introduced. Some Q_B-optimal designs found are then presented, and it is shown that they have achieved advantages in terms of traditional A_S-optimality versus the optimal designs in previous literature for various specified prior probabilities of main effects and two-factor interactions being in the best model.

CO126 **Room K2.31 (Nash Lec. Theatre)** NEW ADVANCES IN SMALL AREA ESTIMATION **Chair: Francesco Schirripa Spagnolo**

C0339: **On the use of small area estimation with geospatial data**
Presenter: **Luciano Perfetti Villa**, University of Southampton, United Kingdom
Co-authors: Nikos Tzavidis, Angela Luna Hernandez

Small-area estimation methods are used for poverty mapping, most commonly with the aid of census data. In the most developed countries, censuses are updated around every ten years but much less frequently than in many countries in the Global South. Geospatial data provide an alternative data source for use in small-area models in off-census years and can improve the frequency of estimates. However, how geospatial data is processed and used in model-based small-area estimation requires careful attention. The purpose is to study theoretically and empirically the properties of small area estimators based on models with geospatial zonal statistics as predictors. The estimators are used to estimate headcount poverty rates for districts in Mozambique. Estimates using geospatial data are compared against estimates produced with the most recent 2017 census (industry standard estimates) and the old 2007 census in Mozambique. The geospatial-based estimates track the industry standard estimates well, but this is not the case for the estimates based on the 2007 census data. The application in Mozambique illustrates the importance of model building/selection when using geospatial data and the potential pitfalls when using old census data.

C1338: **Small area estimation under bivariate Fay-Herriot model with correlated random effects**
Presenter: **Domingo Morales**, University Miguel Hernandez of Elche, Spain
Co-authors: Esteban Cabello, Maria-Dolores Esteban, Agustin Perez Martin

An area-level temporal bivariate linear mixed model is presented, which incorporates correlated time effects for estimating poverty indicators in small areas. The model is applied through the residual maximum likelihood method, leading to the derivation of empirical best linear unbiased predictors for these indicators. Additionally, it provides an approximation of the matrix of mean squared errors (MSE), and it proposes four MSE estimators. The first estimator involves a plug-in approach to the MSE approximation, while the remaining estimators are based on parametric bootstrap procedures. Three simulation experiments were carried out to assess the performance of the fitting algorithm, predictors, and MSE estimators. An application to real data from the 2016 to 2022 Spanish Living Conditions Survey is conducted. The focus is on estimating poverty proportions and gaps for the year 2022, categorized by provinces and sex.

C0457: **Causal small area estimation: The impact of job stability on monetary poverty in Italy**
Presenter: **Setareh Ranjbar**, Lausanne University Hospital , University of Lausanne, Switzerland
Co-authors: Katarzyna Reluga, Nicola Salvati, Dehan Kong, Mark van der Laan

Job stability refers to the security and predictability of employment, including factors such as long-term contracts, adequate wages, social security benefits, and access to training and career development opportunities. Stable employment can play a crucial role in reducing poverty, as it provides individuals and households with a stable income as well as improves their overall and subjective economic well-being. EU-SILC survey and census data are leveraged to assess the causal effect of job stability on monetary poverty across provinces in Italy. To this end, a causal small area estimation (CSAE) framework is proposed for heterogeneous treatment effect estimation in which only a negligible fraction of outcomes is actually observed at the provincial level. The estimators are more stable than the classical causal inference tools as they borrow strength from the other sources of data at the expense of some model assumptions. In a series of model-based and design-based simulations, the influence of different model assumptions on the performance of the proposed algorithm is compared. The new methodology proves to be successful in recovering provincial heterogeneity of the effect of job stability across six regions in Italy.

C0263: **M-quantile regression for zero-inflated data and its applications to small area estimation**
Presenter: **Maria Bugallo**, Miguel Hernandez University of Elche, Spain
Co-authors: Domingo Morales, Francesco Schirripa, Nicola Salvati

Zero-inflated data are almost inevitably complicated by some form of non-observation or inaccurate measurement. From a probabilistic framework, mixtures of GLMMs for the prediction of zero-inflated outcome-dependent indicators have been extensively investigated, and their results are accurate as long as their strong parametric assumptions hold true. However, the demand for results unaffected by outliers in small areas has encouraged the development of new robust inference techniques in recent years. Prompted by the need to develop robust models for variables with an implausible number of zeros, the definition of M-quantiles and their applications are generalized to small area estimation in this field. The contribution includes the proposal of zero-inflated M-quantiles and M-quantile models, the study of asymptotic properties, the derivation of robust predictors, their optimal bias correction and the analytical calculation of mean squared errors. The new methodology is evaluated by means of model-based simulations, showing the gain that the new proposal brings in the presence of just a few atypical data. An application to the Spanish Living Conditions Survey is concluded with.

C0980: **Small area estimation with quantile regression forests**
Presenter: **Nicolas Frink**, Otto-Friedrich-University Bamberg, Germany

A small area estimation method that employs quantile regression forests to enhance the estimation of disaggregated means in small domains is proposed. The use of a machine learning technique allows predictive, non-linear relationships to be captured directly from the data. The approach utilizes quantile-like predictions of the conditional distribution of the outcome variable given the explanatory variables. Thus, analogous to the M-quantile modelling procedure used to estimate means in small areas, the characterization of domain-specific distinctions is facilitated through the generation of domain-specific predictions, thereby overcoming difficulties associated with the identification of random effects. The performance of the quantile regression forests is compared with that of other small area estimation techniques, including both parametric and semi-parametric approaches, using model-based simulations.

CO036 **Room K2.40** HIGH DIMENSIONAL MULTIVARIATE MODELS WITH APPLICATIONS **Chair: Etienne Marceau**

C1386: **Classes of high dimensional Bernoulli distributions and applications**
Presenter: **Patrizia Semeraro**, Politecnico di Torino, Italy
Co-authors: Roberto Fontana

A geometrical structure is provided for classes of high-dimensional Bernoulli distributions that turn out to be important to address open issues in the study of their statistical properties, such as dependence or aggregate risk. The analyzed classes are convex polytopes; in some cases, the extremal generators can be analytically provided; in some other cases, their extremal generators are a more challenging task that can be addressed using an algebraic representation. The class of multivariate Bernoulli distributions are considered to have identical marginal Bernoulli distributions with mean p and the class of multivariate Bernoulli distributions with given sums. The results are applied to study lower bounds for the risk of a credit portfolio.

C1395: **Convex bounds on sums with generalized FGM copula**
Presenter: **Alessandro Mutti**, Politecnico di Torino, Italy
Co-authors: Helene Cossette, Etienne Marceau, Patrizia Semeraro

Building on the one-to-one relationship between generalized FGM copulas and multivariate Bernoulli distributions, it is proven that the class of multivariate distributions with generalized FGM copulas is a convex polytope. Therefore, sharp bounds are found in this class for many aggregate risk measures, such as value-at-risk, expected shortfall, and entropic risk measure, by enumerating their values on the extremal points of the convex polytope. This is infeasible in high dimensions. This limitation is overcome by considering the aggregation of identically distributed risks with generalized FGM copula specified by a common parameter p. In this case, the analogy with the geometrical structure of the class of Bernoulli distribution allows for providing sharp analytical bounds for convex risk measures.

C1434: **Model selection for extremal dependence structures using deep learning: Application to environmental data**
Presenter: **Pierre Ribereau**, Universita Lyon 1, France
Co-authors: Veronique Maume-Deschamps, Manaf Ahmed

The purpose is to introduce a new methodology for extreme spatial dependence structure selection. It is based on deep learning techniques, specifically convolutional neural networks (CNNs). Two schemes are considered: in the first scheme, the matching probability is evaluated through a single CNN, while in the second scheme, a hierarchical procedure is proposed: a first CNN is used to select a max-stable model, then another network allows to select the most adapted covariance function, according to the selected max-stable model. This model selection approach demonstrates good performance on simulations. On the contrary, the composite likelihood information criterion (CLIC) faces issues in selecting the correct model. Both schemes are applied to a dataset of 2m air temperature over Iraq land, CNNs are trained on dependence structures summarized by the concurrence probability.

C1627: **Conditional spatiotemporal copula model for crop insurance**
Presenter: **Melina Mailhot**, Concordia, Canada
Co-authors: Marie Michaelides

Climate change has emerged as one of the most pressing challenges of the time, impacting global ecosystems, economies and social resilience. The insurance industry stands at the forefront of the significantly affected sectors. Accurate prediction of crop yields under varying climatic conditions is paramount for designing sustainable insurance products, determining appropriate premium rates, and ensuring timely payouts that mitigate farmers' financial losses. A spatiotemporal conditional copulas are used, and both ARIMAX-GARCH models and Bayesian regime switching time series are explored for the marginal distributions, which offers a robust approach to risk assessment and premium pricing in agricultural insurance, in addition to providing reliable return level maps. Special cases with closed-form solutions, as well as a comparison between different dependence structures, are presented. An illustration using data from Ontario (Canada) is presented.

C1637: **Coskewness under dependence uncertainty**
Presenter: **Carole Bernard**, Vrije Universiteit Brussel, Belgium
Co-authors: Jinghui Chen, Ludger Ruschendorf, Steven Vanduffel

The impact of dependence uncertainty on $E(X_1X_2...X_d)$ when X_i has cdf F_i for all i is studied. Under some conditions on the F_i, explicit sharp bounds are obtained, and a numerical method is provided to approximate them for arbitrary choices of the F_i. The results are applied to assess the impact of dependence uncertainty on coskewness. In this regard, a novel notion of "standardized rank coskewness" is introduced, which is invariant under strictly increasing transformations and takes values in $[1,1]$.

CO028 Room K2.41 RECENT ADVANCES IN STRUCTURAL EQUATION MODELLING Chair: Andrej Srakar

C1257: **Projection predictive variable selection for Bayesian regularized SEM**
Presenter: **Sara van Erp**, Utrecht University, Netherlands
Co-authors: Paul-Christian Burkner, Aki Vehtari

Regularized structural equation modeling (SEM) provides a useful tool to estimate models with many parameters relative to the sample size without overfitting. An alternative to classical regularization in SEM is Bayesian regularized SEM, in which a shrinkage prior distribution serves as a penalty function. Advantages in terms of flexibility in shrinkage prior choice and automatic uncertainty estimates have resulted in various applications of Bayesian regularized SEM. The goal of Bayesian regularized SEM is often to select a more parsimonious model by including only those parameters in the model which show substantial effects after regularization. Currently, ad-hoc methods are used in SEM to decide if a parameter estimate should be set to zero or not, for example, by relying on an arbitrary threshold value or on the 95% credibility interval. However, it has been shown that the optimal selection criterion depends on various sample and model characteristics. Thus, a formal selection method that works well across different types of SEMs and conditions is needed. A promising method that is available in regression models is projection predictive variable selection, which offers a practical approach to selecting the model that offers nearly similar predictions as a reference model. An extension of the projection predictive method is presented to SEM to determine which parameter estimates are set to zero, thereby performing automatic model selection.

C1327: **A maximum likelihood estimator for composite models**
Presenter: **Tamara Schamberger**, Bielefeld University, Germany
Co-authors: Florian Schuberth, Yves Rosseel, Joerg Henseler

Structural equation modeling (SEM) is a popular and widely applied method that predominantly models latent variables by means of common factor models. Yet, in recent years, the composite model has gained increasing research attention. In contrast to common factor models, approaches to estimate composite models are limited. The contribution is a full-information maximum likelihood (ML) estimator for composite models. The general composite model is presented, a closed form of the variance-covariance matrix implied by this model, and a full information ML estimator is used to obtain the parameter estimates of composite models. Moreover, a test is provided to assess the overall fit of composite models. To demonstrate the performance of the ML estimator and to compare it to its closest contender, i.e., partial least squares path modeling (PLS-PM), in finite samples, a Monte Carlo simulation is conducted. The Monte Carlo simulation reveals that, overall, the ML estimator performs well and is similar to PLS-PM in finite samples. Hence, under the considered conditions, the proposed estimator is a valid alternative with known superior statistical properties.

C1336: **A numerical procedure to estimate dynamic treatment regimes from observational data using structural equation modeling**
Presenter: **Terrence Jorgensen**, University of Amsterdam, Netherlands
Co-authors: Wen Wei Loh

Estimating causal effects from longitudinal observational designs is complicated by natural changes in participation (e.g., students in a multi-year mentoring program to improve academic outcomes may decide to opt-out after better attendance and behavior). Standard comparison of static vs. non-participation has limited real-world relevance, whereas a Dynamic Treatment Regime (DTR) offers greater insight into an interventions efficacy by accounting for differential participation over time. DTRs are an established framework in personalized medicine designed to investigate how individual treatment decisions can optimize intervention effects. Because conducting a randomized study to assess multiple DTRs is rarely feasible practically, how to estimate DTRs using observational longitudinal data is described. Crucially, the estimation procedure resolves the perennial challenges of treatment-dependent confounding inherent in longitudinal designs with treatment-confounder feedback. A numerical procedure is described to estimate causal effects using lavaan, a freely available open-source R package for structural equation modeling, which is widely used in psychological and education sciences. Estimating and interpreting the proposed DTR analysis is demonstrated using educational data as an example. DTRs offer the potential to spur more relevant, tailored, and impactful interventions in psychological and social science research.

C1348: **The structural-after-measurement (SAM) approach to SEM**
Presenter: **Yves Rosseel**, Ghent University, Belgium

In structural equation modeling (SEM), the measurement and structural parts of the model are usually estimated simultaneously. However, since the birth of SEM in the '70s, various authors have advocated that the measurement part should be first estimated, and then the structural part. It is called the structural-after-measurement (SAM) approach. The purpose is to describe the so-called 'local' SAM method, where the mean vector and variance-covariance matrix of the latent variables are expressed as a function of the observed summary statistics and the parameters of the measurement model. The method includes two-step corrected standard errors and local fit measures. Several recent developments are then briefly discussed that are based on the SAM approach, including the inclusion of latent quadratic and interaction terms, the use of non-iterative estimators for the measurement part of the model, small-sample corrections, and various approaches to studying measurement invariance in the setting where the number of groups is very large. Finally, a software implementation of the SAM approach that is available in the R package lavaan is discussed.

C1501: **Improved goodness of fit procedures for structural equation models**
Presenter: **Jonas Moss**, BI Norwegian Business School, Norway
Co-authors: Njaal Foldnes, Steffen Gronneberg

New ways of robustifying goodness-of-fit tests are proposed for structural equation modeling under non-normality. These test statistics have limit distributions characterized by eigenvalues whose estimates are highly unstable and biased in known directions. To take this into account, model-based trend predictions are designed to approximate the population eigenvalues. The new procedures are evaluated in a large-scale simulation study with three confirmatory factor models of varying size (10, 20, or 40 manifest variables) and six non-normal data conditions. The eigenvalues in each simulated dataset are available in a database. Some of the new procedures markedly outperform presently available methods.

CO011 Room S0.11 IMPROVING STATISTICAL IMAGE ANALYSIS Chair: Ranjan Maitra

C0254: **Elliptically-contoured tensor-variate distributions with application to improved image learning**
Presenter: **Carlos Llosa**, Sandia National Laboratories, United States
Co-authors: Ranjan Maitra

Statistical analysis of tensor-valued data has largely used the tensor-variate normal (TVN) distribution, which may be inadequate when data comes from distributions with heavier or lighter tails. A general family of elliptically contoured (EC) tensor-variate distributions is studied, and its characterizations, moments, marginal and conditional distributions are derived. Procedures are described for maximum likelihood estimation from data that are (1) uncorrelated draws from an EC distribution, (2) from a scale mixture of the TVN distribution, and (3) from an underlying but unknown EC distribution, where Tyler's robust estimator is extended. A detailed simulation study highlights the benefits of choosing an EC distribution over the TVN for heavier-tailed data. Tensor-variate classification rules are developed using discriminant analysis and EC errors and show that they better predict cats and dogs from images in the Animal Faces-HQ dataset than the TVN-based rules. A novel tensor-on-tensor regression and tensor-variate analysis of variance (TANOVA) framework under EC errors are also demonstrated to better characterize gender, age and ethnic origin than the usual TVN-based TANOVA in the celebrated Labeled Faces of the Wild dataset.

C0978: **Incorporating seasonality in fMRI time series to address the learning effect**
Presenter: **Israel Almodovar Rivera**, University of Puerto Rico, United States

Functional magnetic resonance imaging (fMRI) is a noninvasive tool for studying regions related to some particular tasks. These activated regions are identified by assigning a map. Reliability across subjects using the activation maps is essential to this analysis. These maps are typically constructed by assigning a general linear model with an autoregressive structure, usually an AR(1) model. However, fMRI experiments naturally have a seasonal structure since the stimulus is applied in each period. A seasonal time series for an fMRI experiment is considered, where the time application of the stimulus is applied. Then, activation maps are obtained from these time series models using classical and adaptive smoothing and thresholding approaches. Reliability was assessed using the Jaccard similarity coefficient as a modified percent activation overlap. The methodology is illustrated in several real datasets from studies involving multiple subjects. It was found that activation maps obtained using seasonal models tend to have more similarities than maps where seasonality was not considered.

C0551: **Improved activation detection from magnitude and phase functional MRI data**
Presenter: **Dan Adrian**, Grand Valley State University, United States
Co-authors: Ranjan Maitra, Daniel Rowe

FMRI data consist of both magnitude and phase components (i.e., it is complex-valued), but in the vast majority of statistical analyses, only the magnitude data is utilized and modeled based on a Gaussian approximation. It is shown that using the correct Ricean distribution for the magnitudes, as well as the entire complex-valued data, results in improved activation detection for activation in the magnitude component. Further, as fMRI measures brain activity indirectly through blood flow, the so-called "brain or vein" problem refers to the difficulty in determining whether measured activation corresponds to (desired) brain tissue or (undesired) large veins, which may be draining blood from neighboring regions. Previous work has demonstrated that activation in the phase component "discriminates" between the two: Phase activation occurs in voxels with large, oriented vessels but not in voxels with small, randomly oriented vessels immediately adjacent to brain tissue. Following this motivation, a model is developed that allows for activation in the phase and magnitude components.

C0687: **Topological data analysis for statistical analysis of structure and dynamics in imaging**
Presenter: **Andrew Thomas**, University of Iowa, United States
Co-authors: Michael Jauch, David Matteson, Peter Crozier

The purpose is to discuss two separate statistical applications of a topological data analysis method introduced called detecTDA for detecting structure and change in a noisy image series. The main image series considered are extremely noisy and highly dynamic catalytic nanoparticle videos from transmission electron microscopy. First, the ability of the topological method to identify structure within the frames of these nanoparticle videos is examined, and whether an image is statistically distinct from one consisting purely of noise is assessed. The method is also demonstrated, along with the newly introduced changepoint method called bclr, pinpoints the statistically significant structural/topological features that govern a change in the state of the nanoparticle video (e.g. from ordered to disordered), concluding with a brief note on the software developed for these tasks.

C0716: **Approximations in the Ising model for use in scene analysis**
Presenter: **Alejandro Murua**, University of Montreal, Canada
Co-authors: Ranjan Maitra

The Ising distribution is useful in many applications involving statistical modelling and inference. However, its normalization is constant, and its moments are intractable. This drawback has prevented a wider use of this model. Analytical approximations are provided that make it possible to estimate these quantities numerically in a homogeneous case. Simulation studies indicate good performance compared to Markov chain Monte Carlo methods and in a tiny fraction of the time. The methodology is illustrated in performing Bayesian inference in a functional magnetic resonance imaging activation detection experiment.

CO321 Room Safra Lec. Theatre SNAPSHOT ON CURRENT FUNCTIONAL DATA METHODOLOGIES Chair: Frederic Ferraty

C0218: **Empowering multi-class classification for complex functional data with simultaneous feature selection**
Presenter: **Guanqun Cao**, Michigan State University, United States

The opportunity to utilize complex functional data types for conducting classification tasks is emerging with the growing availability of imaging data. However, the tools capable of effectively managing imaging data are limited, let alone those that can further leverage other one-dimensional functional data. Inspired by the extensive data provided by the Alzheimer's Disease Neuroimaging Initiative (ADNI), a novel classifier is introduced tailored for complex functional data. Each observation in this framework may be associated with numerous functional processes varying in dimensions, such as curves and images. Each predictor is a random element in an infinite dimensional function space, and the number of functional predictors p can potentially be much greater than the sample size n. In contrast to the existing functional data classifiers, the proposed unified model performs feature selection and classification simultaneously. The challenge arises from the complex inter-correlation structures among multiple functional processes and without any assumptions on the distribution of these processes. A simulation study and real data application are carried out to demonstrate its favorable performance.

C0230: **Factor-augmented functional regression with an application to electricity price curve forecasting**
Presenter: **Sven Otto**, University of Cologne, Germany
Co-authors: Luis Winter

A function-on-function linear regression model is proposed for time-dependent curve data that is consistently estimated by imposing factor structures on the regressors. A novel integral operator based on cross-covariances identifies two components for each functional regressor: a predictive low-dimensional component, along with associated factors that are guaranteed to be correlated with the dependent variable, and an infinite-dimensional component that has no predictive power. In order to consistently estimate the correct number of factors for each regressor, a functional eigenvalue difference test is introduced. The model is applied to forecast electricity price curves in three different energy markets. Its prediction accuracy is found to be comparable to popular machine learning approaches while providing interpretable insights into the conditional correlation structures of electricity prices.

C0617: **Penalized regression models informed by physics**
Presenter: **Michelle Carey**, Univerity College Dublin, Ireland
Co-authors: James Ramsay

Linear partial differential equations (PDEs) are powerful tools for modeling complex spatial autocorrelation in irregularly shaped domains. However, estimating parameters for these equations is a challenging task. To address this, a penalized regression framework is introduced. This method uses partially observed and noisy data to estimate PDE parameters and quantify the uncertainty of these estimates and the resulting PDE solution. Simulations demonstrate that this approach significantly outperforms existing methods, achieving a three-fold improvement in parameter estimation accuracy. The method is also applied to Croatian temperature data to showcase its effectiveness in spatial data analysis. By leveraging the PDE, estimates of the spatial process are generated in complex domains with irregular data sampling and intricate spatial dependencies.

C1067: **On advances in shape-based functional data analysis**
Presenter: **Anuj Srivastava**, Florida State University, United States

Functional data analysis (FDA) is a fast-growing area of research and development in statistics. While most FDA literature imposes the classical l_2 Hilbert structure on function spaces, there is an emergent need for a different, shape-based approach for analyzing functional data. Fundamental geometrical concepts are reviewed and developed to help connect traditionally diverse fields of shape and functional analysis. It showcases that focusing on shapes is often more appropriate when structural features (number of peaks and valleys and their heights) carry salient information in data. It recaps recent mathematical representations and associated procedures for comparing, summarizing, and testing the shapes of functions. Specifically, it discusses shape regression models that extract and focus on the shapes of functions during regression. The ensuing results provide better interpretations and tend to preserve geometric structures.

C1345: **Fast rate estimation of integrals of multivariate random functions**
Presenter: **Sunny Wang**, ENSAI, France
Co-authors: Valentin Patilea

The problem of estimating integrals of random functions defined on a compact interval or multidimensional rectangle arises naturally in functional data analysis. Such integrals are needed to make predictions using functional regression models, to compute the scores of a random function in a basis, to compute data depths, etc. Using a recent linear integration approach from the Monte Carlo literature, a new class of estimators is proposed in the case where the random functions are observed, possibly with measurement noise, on a discrete, random set of design points. In the absence of noise, the estimators converge faster than the empirical means. Narrow confidence intervals are proposed both with and without measurement noise.

CO246 Room BH (S) 1.01 Lec. Theathre 1 RECENT CONTRIBUTIONS IN APPLIED ECONOMETRICS Chair: Pipat Wongsa-art

C0541: **The impact of Brexit on ethnic discrimination among NHS Workforce**
Presenter: **Victoria Serra-Sastre**, City, University of London, United Kingdom
Co-authors: Catia Nicodemo

The effects of Brexit on experiences of workplace discrimination are examined among clinical staff from ethnic minority backgrounds working in the UK's National Health Service (NHS). Using difference-in-differences models and data from the NHS Staff Survey over 2012-2022, evidence is provided that minority doctors, nurses, midwives and healthcare assistants reported increased discrimination from both patients/public and peers/managers after the 2016 Brexit referendum relative to trends in discrimination facing white staff. These findings point to concrete human costs of Brexit for minority NHS personnel that have received little attention to date despite the health service's heavy reliance on minority and immigrant workers. Monitoring staff experiences and supporting diversity must be priorities for retaining the NHS's diverse workforce in the post-Brexit era.

C0614: **Almost unbiased variance estimation in IV regression**
Presenter: **Yongdeng Xu**, Cardiff University, United Kingdom

Using Nagar's approximation, bias approximations are derived for the asymptotic variance estimators of 2SLS and Fuller estimators in instrumental variable regression, up to the order of T^{-2}. These bias approximations are then employed to develop bias-corrected variance estimators. Findings indicate that the asymptotic variance estimator for 2SLS exhibits an upward bias, while the Fuller estimator shows an upward bias in cases of just identification or low overidentification. Simulation results reveal that the bias-corrected variance estimator substantially reduces the bias of the asymptotic variance estimator, demonstrating minimal bias or near-unbiasedness. Importantly, the properties of the t-test improve significantly with bias-corrected variance estimates, consistently demonstrating better size properties and higher statistical power compared to the asymptotic variance. The practical relevance of these findings is illustrated through several well-known applications.

C0656: **The Nash wage elasticity and its business cycle implications**
Presenter: **Matthew Knowles**, City, University of London, United Kingdom

A new measure of wage rigidity, the Nash wage elasticity (NWE), is developed and used to evaluate the importance of wage rigidity for business cycles. The NWE is the percentage change in the actual wage rate when the wage that would occur under Nash bargaining changes by 1%. It is shown that the NWE can be measured from aggregate data under relatively weak assumptions which hold across a large class of search and matching models. The empirical value can then be compared with the values predicted by specific models in this class. In the US data, the estimates of the NWE are generally between 0 and 0.1, indicating that, for both continuing workers and new hires, (a) there is a high degree of wage rigidity, and (b) Nash bargaining provides a poor description of wage setting. It is shown that the estimates imply that wage rigidity greatly amplifies business cycles: A simple SAM model suggests that if workers were paid Nash-bargained wages rather than actual wages, then the cyclical volatility of unemployment would decrease to less than one-seventh of what it is in the data. The results are compared to various models of rigid and flexible wages in the literature.

C0920: **Parametric Whittle estimation of cyclically integrated time series**
Presenter: **Edward Hill**, Queen Mary University of London, United Kingdom
Co-authors: Liudas Giraitis

Macroeconomic and climatic time series may contain unknown or hidden periodicities characterized by a singularity in the spectral density function at a non-zero frequency. Recently, a study introduced a parametric ARCIMA model for the modelling of cyclically integrated time series. The focus is on the estimation of parametric ARCIMA models. The inference is performed in the frequency domain using a modified Whittle likelihood function, where the location of the singularity in the spectral density has been estimated using a separate procedure developed in a recent study. The parametric rate of consistency is derived, and the estimator's asymptotic normality is established. Simulations confirm the good performance of the two-stage estimation procedure in finite samples.

C0749: **A new model for agricultural land use modelling and prediction in England using spatially high-resolution data**
Presenter: **Pipat Wongsa-art**, City St Georgeś, University of London, United Kingdom

The focus is on better understanding farmers' responses to behavioural drivers of land-use decisions by establishing an alternative analytical procedure that can handle complex data structures and overcome various methodological drawbacks suffered by methods currently used in existing studies. Firstly, the procedure uses spatially high-resolution data so that idiosyncratic effects of physical environment drivers, e.g. soil textures, can be explicitly modelled. Secondly, the well-known censored data problem is addressed, which often hinders a successful analysis of land-use shares. Thirdly, spatial error dependence and heterogeneity are incorporated in order to obtain efficiency gain and a more accurate formulation of variances for the parameter estimates. Finally, the computational burden is reduced, and estimation accuracy is improved by introducing an alternative GMM-QML hybrid estimation procedure. The newly proposed procedure is applied to spatially high-resolution data in England and found that, by taking these features into consideration, conclusions are formulated about the causal effects of climatic and physical environment, and environmental policy on land-use shares that differ significantly from those made based on methods that are currently used in the literature. Moreover, it is shown that the method enables the derivation of a more effective predictor of land-use shares, which is utterly useful from the policy-making point of view.

CO069 **Room BH (SE) 1.01** ADVANCES IN HIGH/INFINITE-DIMENSIONAL INFERENCE (VIRTUAL) **Chair: Catia Scricciolo**

C0894: **Polynomial time guarantees for sampling based posterior inference**
Presenter: **Randolf Altmeyer**, Imperial College London, United Kingdom

The Bayesian approach provides a flexible framework for a wide range of non-parametric inference problems. It relies crucially on computing functionals with respect to the posterior distribution, such as the posterior mean or posterior quantiles for uncertainty quantification. Since the posterior is rarely available in closed form, this is based on Markov chain Monte Carlo (MCMC) sampling algorithms. The runtime of these algorithms, until a given target precision is achieved, will typically scale exponentially in the model dimension and the sample size. In contrast, sampling-based posterior inference in a general high-dimensional setup is shown as feasible, even without global structural assumptions such as strong log-concavity of the posterior. Given a sufficiently good initializer, polynomial-time convergence guarantees are presented for a widely used gradient-based MCMC sampling scheme. The key idea is to combine posterior contraction with the local curvature induced by the Fisher-information of the statistical model near the data-generating truth. Applications to high-dimensional logistic and Gaussian regression are discussed.

C0918: **Recent advances in high-dimensional Bayesian graphical models**
Presenter: **Sayantan Banerjee**, Indian Institute of Management Indore, India

Advances in technology have resulted in massive datasets collected from all aspects of modern life. Very large datasets appear from internet searches, mobile apps, social networking, cloud computing, and wearable devices, as well as from more traditional sources such as bar-code scanning, satellite imaging, air traffic control, banking, finance, and genomics. Due to the complexity of such datasets, flexible models are required, which often involve many parameters that routinely exceed the available sample size. In such a situation, a meaningful inference is possible only if there is a hidden lower-dimensional structure involving far fewer parameters; that is, the system is sparse. The problem of Bayesian estimation of a high-dimensional precision (inverse covariance) matrix corresponding to a Gaussian graphical model under continuous shrinkage priors is considered. Some recent theoretical and computational advances in this field are discussed. Computationally efficient Markov chain Monte Carlo methods and expectation conditional maximization algorithms are provided, respectively, for fully Bayesian and penalized likelihood problems. They also provide theoretical guarantees of the related methods, including establishing posterior convergence. The approach is also validated through extensive simulation studies and via applications in bioinformatics, psychology, finance, and genomics.

C0286: **Bayesian fixed-domain posterior contraction for spatial Gaussian process model with nugget**
Presenter: **Cheng Li**, National University of Singapore, Singapore
Co-authors: Saifei Sun, Yichen Zhu

Spatial Gaussian process regression models typically contain finite dimensional covariance parameters that need to be estimated from the data. The Bayesian estimation of covariance parameters is studied, including the nugget parameter in a general class of stationary covariance functions under fixed-domain asymptotics, which is theoretically challenging due to the increasingly strong dependence among spatial observations. A novel adaptation of Schwartz's consistency theorem is proposed for showing posterior contraction rates of the covariance parameters, including the nugget. A new polynomial evidence lower bound is derived, and consistent higher-order quadratic variation estimators are proposed that satisfy concentration inequalities with exponentially small tails. The Bayesian fixed-domain asymptotics theory leads to explicit posterior contraction rates for the microergodic and nugget parameters in the isotropic Matern covariance function under a general stratified sampling design. The theory and the Bayesian predictive performance are verified in simulation studies and an application to sea surface temperature data.

C0183: **On the minimax rate of the Gaussian sequence model under (bounded) convex constraints**
Presenter: **Matey Neykov**, Northwestern University, United States

The purpose is to discuss the exact minimax rate of a Gaussian sequence model under (bounded) convex constraints, purely in terms of the local geometry of the given constraint set. It is argued that the local entropy of the set K determines the rate via a certain fixed point equation. The results are extendable to nonparametric density estimation as well as nonparametric regression.

C0522: **Variational Bayesian procedures with frequentist guarantees**
Presenter: **Dennis Nieman**, VU Amsterdam, Netherlands

Among the various reasons for the use of variational Bayesian (VB) methods are the possibility of model selection and the reduction of computation time. A theoretical approach to these motivations is taken, studying variational posteriors from a frequentist perspective. Using minimax theory, statistical and computational needs are balanced by finding the minimal dimension of the VB approximation to the posterior required to achieve the optimal convergence rate. The coverage of variational posterior credible regions is also studied. VB model selection entails the tuning of hyperparameters by optimization of the evidence lower bound, which can lead to smoothness-adaptive convergence rates.

CO180 Room BH (SE) 1.02 ADVANCES IN BAYESIAN MIXTURE MODELING APPLIED TO COMPLEX DATASETS Chair: Andrea Sottosanti

C0353: **Bayesian mixture of spectral density functions for ocean waves**
Presenter: **Matt Moores**, University of Wollongong, Australia
Co-authors: Sankalpa Fonseka, Jeff Hansen, David Gunawan

Ocean waves are a fundamental part of the environment and play an important role in the physical and biological processes of the oceans. The distribution of energy versus wave frequency, known as a wave spectrum, provides a way to quantify the behaviour of ocean waves. Accurate estimation of the physical parameters, such as wave height, peak frequency, and velocity, is crucial for a wide range of applications, including offshore engineering, coastal management, and maritime operations. A mixture of spectral density functions is introduced to model ocean wave spectra with multiple peaks, i.e., wind seas and swell. Informative priors for the parameters are defined using expert elicitation. Posterior samples are obtained using Hamiltonian Monte Carlo. Simulation-based calibration is employed to quantify the accuracy, consistency and coverage of the estimates. The model is fit to ocean buoy measurements from King George Sound in Western Australia.

C0630: **Bayesian mixture models for scientific discovery in astrophysics**
Presenter: **David van Dyk**, Imperial College London, United Kingdom

Mixture data are ubiquitous in astrophysics. Photons originating from different sources or from different physical processes within a single source are often indistinguishable. Point sources (such as stars) may be physically embedded in an extended source (such as nebulae), or sources may simply appear to overlap with their foreground or background from the vantage point of Earth. Astrophysicists inevitably aim to use photons from each source to understand its individual physical properties but only have a mixture of photons from all sources. Even with well-separated sources, photons originate from multiple physical processes. A star that appears as a point source, for example, may contain both turbulent/flaring regions and quiescent regions, and even in a single region, the star's plasma may exhibit multiple physical processes with different but mixed spectral signatures. A population of sources may contain subclasses that are difficult to distinguish, yet researchers wish to analyze them separately. Instrumental effects such as point-spread functions, photon-redistribution matrices, and pile-up also effectively mix photons. Several specific examples are discussed where the creative deployment of Bayesian mixture models in astronomy has enabled overlapping sources/processes to be disentangled, scientific parameters of individual sources to be estimated, and uncertainty to be properly quantified.

C0693: **Bayesian co-clustering of ordinal data with informative censoring**
Presenter: **Alice Giampino**, University of Milano-Bicocca, Italy
Co-authors: Antonio Canale, Bernardo Nipoti

A novel Bayesian nonparametric model is proposed for co-clustering multivariate ordinal data. The ordinal nature of the data is addressed through a latent variable framework, while its large dimensionality is managed via a matrix factorization model. The flexibility of the approach stems from the use of two independent Dirichlet processes, which allow for inference on the number of clusters within the latent structure. Unlike most approaches available in the literature, the method treats censored observations as potentially informative rather than absent information, reflecting that missing information can be valuable in profiling individual preferences. Thanks to the conjugate specification of the model, which allows for the explicit derivation of the full conditional distributions, posterior inference is performed using a Gibbs sampling algorithm. The performance of the method is demonstrated using data on politician votes.

C0762: **Flexible modeling of grouped multivariate data via Bayesian shared-atom nested mixture models**
Presenter: **Laura D Angelo**, Universita di Milano Bicocca, Italy
Co-authors: Federica Zoe Ricci

The use of hierarchical mixture priors with shared atoms has recently flourished in the Bayesian literature for partially exchangeable data. Leveraging on nested levels of discrete random measures, these models allow the estimation of a two-layered data partition: across groups and among observations. We illustrate the properties of such modeling strategies when the mixing weights are assigned either a finite-dimensional Dirichlet distribution or a Dirichlet process prior. We discuss the use of hierarchical nonparametric priors based on a finite set of shared atoms, showing how this specification preserves the flexibility of the induced random measure and allows the derivation of fast posterior inference. Specifically, we develop a mean-field variational algorithm for posterior inference to boost the applicability to large multivariate data. This allows us to fit a nested mixture model to a real dataset of Spotifys song features, simultaneously segmenting artists and songs with similar characteristics.

C0778: **Combining Bayesian latent traits and topic models for identifying risky behavioral profiles in Italian population**
Presenter: **Mattia Stival**, Ca Foscari University of Venice, Italy
Co-authors: Angela Andreella, Lorenzo Schiavon, Stefano Campostrini

Understanding the relationships between risk factors is vital in health economics and policy making, facilitating targeted campaigns for vulnerable population subgroups. Behavioral and risk factors surveillance systems (BRFSS) provide extensive data, including demographics, socio-economics, and behavioral habits. Among these, employment information is collected as textual data. Despite its relevance, linking employment information with other risk factors is challenging, and few attempts have been observed in the literature. A model is introduced that integrates latent Dirichlet allocation (LDA) for textual analysis with latent trait models (LTM) for categorical data. LDA identifies topics that describe unobservable job macro-groups. These macro-groups are likely characterized by heterogeneous propensities to risky behavior, aspects captured by the LTM. To better describe this variability, covariates are included in the LTM and let it depend on the topics identified by LDA. A modular inferential procedure is proposed that uses the state of-the-art methods from both approaches, combining their results using importance sampling. The inferential procedure leads to a valid posterior inference strategy, fully leveraging the methodological advancements of both models without the need to develop new cumbersome algorithms from scratch.

CO017 Room BH (SE) 1.05 WEATHERING THE CLIMATE STORM: FINANCIAL RESILIENCE Chair: Juan-Angel Jimenez-Martin

C1355: **Assessing the ESG premium: Evidence from Spain**
Presenter: **Simon Sosvilla-Rivero**, Universidad Complutense de Madrid, Spain
Co-authors: Julian Andrada-Felix , Marta Gomez-Puig

As a case study, the premium investors assign after a company is included in the FTSE4Good IBEX Index is assessed, an index designed to identify

Spanish companies with leading corporate responsibility practices. By comparing the performance of companies included in the index with that of a counterfactual company with similar characteristics but has not been selected, it is quantified how market participants have valued the strong management of environmental, social, and governance risks by the company under study. The simultaneous nearest neighbors approach is used to create the counterfactual company by simultaneously considering occurring analogues in the time series of other companies not selected and listed in the Madrid Stock Market.

C1443: **Measuring the impact of biodiversity loss on financial markets: A CoVaR approach**
Presenter: **Lidia Sanchis-Marco**, University of Castilla-La Mancha, Spain
Co-authors: Laura Garcia-Jorcano

Combining biodiversity data is conducted with sector and system returns, a quantile regression analysis of the impact of physical and transition risks associated with biodiversity loss on extreme sector profitability and losses, and system losses, as well as the impact of sector losses and profits on biodiversity loss. To this end, a world biodiversity index is constructed, and a risk management measure is used to capture its extreme loss values over time. Based on a systemic risk measure by another study, a novel approach is proposed to measure financial risk conditional on biodiversity loss. This method allows the creation of a financial risk measure called CoBiodiversity that captures the dependence of extreme losses and profits of different sectors and system losses on changes in biodiversity loss at extreme quantiles and vice versa. Related market measures are also introduced called DeltaCoBiodiversity for each sector and the system exposure to biodiversity loss and ExposureCoBiodoversity to assess the impact of sector profits and losses on biodiversity loss. The main evidence indicates the highest CoBiodiversity and DeltaCoBiodiversity measures for system losses and sector profits conditioned to biodiversity loss, especially in information technology and financials, during the global financial crisis of 2008 and COVID-19. The highest ExposureCoBiodiversity measure is also found for sector profits for materials and energy sectors in 2008 and 2020, respectively.

C1512: **A tale of dynamic tail climate transition risk exposure: TCaRE**
Presenter: **Juan-Angel Jimenez-Martin**, Complutense of Madrid, Spain
Co-authors: Laura Garcia-Jorcano, M Dolores Robles

TCaRE is introduced, which measures the exposure of industry portfolios to climate transition risks (CTR). CTR, a significant challenge for firms transitioning to a low-carbon economy, negatively impacts corporate profitability. The analysis reveals that TCaRE is influenced by both firm-specific characteristics and broader market trends. Industry dynamics, climate events, and external factors such as market volatility and policy uncertainty contribute to the complexity of TCaRE. Understanding TCaRE is crucial for investors and managers in navigating the evolving landscape of climate-related financial risks.

C1504: **Does the gender diversity in the nexus with sustainability affect downside and tail risks**
Presenter: **Almudena Maria Garcia Sanz**, Complutense University of Madrid, Spain
Co-authors: Juan-Angel Jimenez-Martin, M Dolores Robles

The role of diversity on boards of directors is examined in the nexus between sustainability (ESG) practices and the downside and tail risks of firms between 2010 and 2022. A set of US and European companies included in the Refinitiv DataStream indexes are analyzed. The sample period spans from 2010 to 2022. A panel regression analysis is applied to explain a comprehensive set of firms' downside and tail risk dimensions (tail beta, downside beta, lower partial moment, value at risk and expected shortfall) as a function of their commitment to sustainability practices as proxied by Thomson Reuters ESG scores and their commitment to gender diversity as a proxy by a set of measures of gender and cultural diversity. Results indicate the relevance of ESG practices and the diversity of the boards in risk management, underscoring the relevance of ESG practices and board diversity in mitigating risks, even after controlling for firm characteristics such as size and profitability and board characteristics as board size and board tenure. The analysis is split to focus separately on each ESG pillar to find the driver or drivers of the risk mitigation in the nexus with each of the diversity measures. US and European companies are also analyzed separately, and significant differences in the impact of ESG practices are observed on each risk dimension to each of the diversity measures. In-depth analysis is undertaken to focus on the behavior by industries.

C1568: **Financing biodiversity: Does a biodiversity premium exist in sustainable bonds**
Presenter: **Enrique Ballesta**, URJC, Spain
Co-authors: Pilar Abad, M-Dolores Robles

Biodiversity is a critical component of ecosystem services, providing environmental sustainability and economic value. Even so, biodiversity has suffered unprecedented damage during the last decades, and its degradation poses a major risk to the global economy and the financial system. In that context, sustainable debt has gained the attention of investors aiming to build nature risk-resilient portfolios, and issuers, financial institutions and regulators have increased its awareness. The objective aims to tackle the biodiversity finance research gap, specifically focusing on European sustainable debt, by studying the existence of a biodiversity yield premium, biodiversity greenium in biodiversity-oriented sustainable bonds. To do so, econometric methods are proposed, specifically pseudo-panel data models, fed with public market data obtained by using an unprecedented matching method between the sustainable bond and a brown quoted curve. A database has been developed incorporating European sustainable labelled debt issued during the period between 2007 and 2022 with information to study the existence of any geographical influence or temporal influence in the biodiversity greenium in the periods around the UN COPs meetings. The dataset also incorporates unprecedented information about the proceeds' alignment with the preservation of biodiversity and with respect to the UN SDGs.

CO159 Room BH (SE) 1.06 ADVANCES IN BAYESIAN METHODS Chair: Lucia Paci

C0440: **Bayesian nonparametric mixtures of categorical directed graphs for heterogeneous causal inference**
Presenter: **Federico Castelletti**, Universita Cattolica del Sacro Cuore (Milan), Italy
Co-authors: Laura Ferrini

Quantifying the causal effects of exposures on outcomes, such as a treatment and a disease, respectively, is a crucial issue in medical science for the administration of effective therapies. Importantly, any related causal analysis should account for all those variables, e.g. clinical features, that can act as risk factors involved in the occurrence of a disease. In addition, the selection of targeted strategies for therapy administration requires quantifying such treatment effects at a personalized level rather than at a population level. These issues are addressed by proposing a methodology based on categorical directed acyclic graphs (DAGs), which provide an effective tool for inferring causal relationships and causal effects between variables. In addition, population heterogeneity is accounted for by considering a Dirichlet process mixture of categorical DAGs, which clusters individuals into homogeneous groups characterized by common causal structures, dependence parameters and causal effects. Computational strategies are developed for Bayesian posterior inference, from which a battery of causal effects at the subject-specific level is recovered. The methodology is evaluated through simulations and applied to a dataset of breast cancer patients to investigate side effects associated with the occurrence of cardiotoxicity and possibly implied by the administration of anticancer therapies.

C0561: **Bayesian structural learning with parametric marginals for count data: An application to microbiota systems**
Presenter: **Pariya Behrouzi**, Wageningen University and Research, Netherlands
Co-authors: Veronica Vinciotti, Reza Mohammadi

High-dimensional and heterogeneous count data are collected in various applied fields. The aim is to look closely at high-resolution sequencing data on the microbiome, which has enabled researchers to study the genomes of entire microbial communities. Revealing the underlying interactions between these communities is of vital importance to learn how microbes influence human health. To perform structural learning from multivariate count data such as these, a novel Gaussian copula graphical model is developed with two key elements. Firstly, parametric regression is employed to characterize the marginal distributions. This step is crucial for accommodating the impact of external covariates. Neglecting this adjustment could potentially introduce distortions in the inference of the underlying network of dependences. Secondly, a Bayesian structure learning framework is advanced, based on a computationally efficient search algorithm that is suited to high dimensionality. The approach returns simultaneous inference of the marginal effects and of the dependence structure, including graph uncertainty estimates. A simulation study and a real data analysis of microbiome data highlight the applicability of the proposed approach in inferring networks from multivariate count data in general and its relevance to microbiome analyses in particular. The proposed method is implemented in the R package BDgraph.

C1230: **Dependent Dirichlet processes via thinning**
Presenter: **Bernardo Nipoti**, University of Milan Bicocca, Italy
Co-authors: Laura D Angelo, Andrea Ongaro

An easy-to-implement strategy is proposed to define a flexible class of dependent Dirichlet processes using a thinning technique. Specifically, the well-known stick-breaking construction of the Dirichlet process is modified by introducing random breaks that follow a mixture distribution: a point mass at zero (indicating no break) and a Beta distribution (as in the Dirichlet process). This modification results in the definition of dependent processes that retain analytical tractability, allowing the thorough study of their properties and efficient algorithm development for posterior computation. The approach is illustrated through its application to a two-level clustering problem in the analysis of the Collaborative Perinatal Project data, showcasing its ability to cluster both patients and hospitals simultaneously.

C0186: **Consistent and fast inference in compartmental models of epidemics using Poisson approximate likelihoods**
Presenter: **Lorenzo Rimella**, University of Turin, Italy
Co-authors: Michael Whitehouse, Nick Whiteley

Addressing the challenge of scaling up epidemiological inference to complex and heterogeneous models, Poisson approximate likelihood (PAL) methods are introduced. In contrast to the popular ordinary differential equation (ODE) approach to compartmental modelling, in which a large population limit is used to motivate a deterministic model, PALs are derived from approximate filtering equations for finite-population, stochastic compartmental models, and the large population limit drives consistency of maximum PAL estimators. Theoretical results appear to be the first likelihood-based parameter estimation consistency results which apply to a broad class of partially observed stochastic compartmental models and address the large population limit. PALs are simple to implement, involving only elementary arithmetic operations and no tuning parameters, and fast to evaluate, requiring no simulation from the model and having computational cost independent of population size. Through examples, it is demonstrated how PALs can be used to: Fit an age-structured model of influenza, taking advantage of automatic differentiation in Stan; compare over-dispersion mechanisms in a model of rotavirus by embedding PALs within sequential Monte Carlo; and evaluate the role of unit-specific parameters in a meta-population model of measles.

CO271 Room BH (S) 2.01 MODELING FINANCIAL TIME SERIES WITH CONOMETRICS AND MACHINE LEARNING Chair: Markus Haas

C0322: **A new way to specify dynamic models**
Presenter: **Leopoldo Catania**, Aarhus BBS, Denmark

A new general class of models, which lies between observation-driven and parameter-driven models, is presented. Several examples are discussed in the context of conditional mean and variance specifications. Conditions for strong stationarity and the existence of moments are derived. Consistency and asymptotic normality of the maximum likelihood estimator are derived for the general model specification and for specific examples. An application in financial econometrics shows the usefulness of the proposed class of models.

C0739: **Good volatility, bad volatility and time-varying skewness**
Presenter: **Dennis Umlandt**, University of Innsbruck, Austria

A parametric model of good and bad volatility is proposed and studied with time-varying higher-order moments. Volatilities follow observation-driven updating schemes to minimize the conditional second and third-order moment criterion. As a result, gamma-distributed good and bad shocks are identified by their effect on the skewness of the series rather than strictly by their sign, as in typical asymmetric volatility models. Estimation and inference are performed using straightforward likelihood maximization. Monte Carlo evidence suggests that the novel approach is able to recover heterogeneous dynamics in good and bad volatility. The model is applied empirically to US stock returns, and it is found that their distribution is more negatively skewed after adverse shocks and that the skewness dynamics can explain the asymmetric responses of volatility.

C0817: **Realized volatility forecasting for new issues and spin-offs using multi-source transfer learning**
Presenter: **Andreas Teller**, Friedrich Schiller University Jena, Germany
Co-authors: Uta Pigorsch, Christian Pigorsch

The special case of forecasting realized volatility of financial assets with limited historical data is considered, such as new issues or spin-offs. Typically, realized volatility forecasting models rely on a sufficient history of data. For new issues and spin-offs, however, an extensive data history is not directly available. Therefore, the proposal is to forecast the realized variance of assets with limited historical data based on multi-source transfer learning. Specifically, complementary source data of financial assets is exploited with a substantial historical data record by selecting source time series instances most similar to the target data of the respective new issue or spin-off. Based on these instances and the target data, heterogeneous autoregressive models, feedforward neural networks, and extreme gradient boosting models are estimated. Their forecasting performance is compared to forecasts of the same models trained exclusively on the target data and to a simplified training data pooling approach that includes the entire source and target data. Results indicate that integrating complementary data can significantly improve the accuracy of realized variance forecasts for new issues and spin-offs, even shortly after their initial trading day. In particular, the proposed transfer learning approach shows superior performance compared to models trained solely on target asset data and those that additionally incorporate the complete source data.

C0678: **Supply chain disruptions and foreign exchange rate volatility**
Presenter: **Mawuli Segnon**, University of Munster, Germany

The relationship between supply chain disruptions and foreign exchange (FX) rate volatility is investigated in fundamentals-based component frameworks. MIDAS volatility models are utilized to formalize FX rate volatility as a product of short- and long-run components. The long-run components are allowed to be driven by the monthly global supply chain pressure index (GSCPI) and the global economic condition (GECON) indicator, as well as some economic fundamentals such as inflation rates, policy rates, money differentials and industrial production. The dynamics governing the short-run component are modeled via GARCH, GJR, Markov switching multifractal (MSM) and factorial hidden Markov volatility (FHMV) processes. Empirical application results show a statistically significant positive relationship between supply chain disruptions and foreign exchange rate volatility for developed and developing countries. The predictive content of the global supply chain pressure index is evaluated based on up-to-date model comparison and backtesting tests. The results show a substantial improvement in the accuracy of out-of-sample forecasts of future FX volatility and value-at-risk for eleven developed and developing countries.

C0935: **A multiple chains hidden Markov model for a sector index and its comovement with the market**
Presenter: **Markus Haas**, University of Kiel, Germany

Bivariate multiple chains hidden Markov model is considered and applied to characterize the regime-switching dynamics of the US airline industry and its comovement with the S&P 500 during the Covid-19 pandemic. The model allows for independent regime switches of the sector-specific and the market factors, with the prevailing states of both chains determining the parameters of the conditional sector return distribution. Shocks are drawn from a (possibly regime-specific) Student's t distribution to capture the fat-tailed nature of the returns and to avoid the otherwise distorting effect of outliers on the identification of the regime processes. An EM-type algorithm is proposed to estimate the model parameters.

CO205 Room BH (S) 2.02 RECENT ADVANCES IN WELL-BEING AND POVERTY MEASUREMENT Chair: Chiara Gigliarano

C0313: **A new point of view in the construction of composite indicators: A case study**
Presenter: **Francesca Mariani**, Universita Politecnica delle Marche, Italy
Co-authors: Maria Cristina Recchioni, Mariateresa Ciommi, Chiara Gigliarano

The construction of composite indicators (CI) is a challenging operation. A new technique is proposed to construct composite indicators that are based on the Box-Cox transformation. To illustrate the idea, a case study has been carried out. Results show that it is possible to find a parameter from the Box-Cox transformation that, in the least squares sense, ensures normality and, as a consequence, opens new inroads for researchers in the field of the construction of CI.

C0514: **Regional comparison of household well-being: Insights from Lombardy, Tuscany, and Campania**
Presenter: **Antonella D Agostino**, University of Siena, Italy
Co-authors: Laura Neri, Andrea Regoli, Gianni Betti, Fernando Tavares

The purpose is to offer a comparative analysis of the living conditions of households in three different regions of Italy: Lombardy, Tuscany and Campania. Each of these regions exemplifies the different quality of life and living conditions typical of the North, Centre and South of the country, respectively. The research makes use of data from a new sample survey conducted in 2024. 2024, as part of the PRIN 2022 PNRR project entitled "MYPEOPLE: Measuring inequality, poverty and living conditions for local strategy planning". The primary objective is to assess and compare the vulnerability of households in these regions using a multidimensional and fuzzy approach. This approach allows for the analysis of multiple dimensions of vulnerability, including economic stability, working conditions, education and social inclusion, thus providing a complete picture of living conditions in Lombardy, Tuscany and Campania. The fuzzy aspect of the framework helps manage the inherent uncertainty of the phenomenon under investigation, providing a more nuanced understanding of household well-being. The results are expected to offer valuable insights for policymakers and practitioners involved in planning and implementing local strategies that address specific dimensions of vulnerability to improve living conditions and reduce inequalities.

C0874: **Dynamic approaches to ranking happiness: Integrating benefit of the doubt weighting and functional data analysis**
Presenter: **Annalina Sarra**, University of Chieti-Pescara, Italy
Co-authors: Eugenia Nissi, Adelia Evangelista, Tonio Di Battista

In recent years, the pursuit of happiness has become a key indicator of societal well-being, prompting many countries to measure their Happiness Index. Traditional rankings use static data, which may not fully capture the multifaceted nature of happiness over time. A novel approach is introduced, combining benefit of the doubt (BoD), weighting with functional data analysis (FDA) to create a comprehensive Happiness Index ranking. BoD, used in composite indicators (CIs), flexibly derives weights from the data, allowing for the aggregation of quantitative sub-indicators without precise weight information. After establishing a CI of happiness, the FDA is used to examine how happiness scores evolve over time. Two functional measures are specifically utilized: the modified hypograph index (MHI) and the weighted integrated first derivative (DW). The MHI assesses the proportion of time during which one country's performance surpasses others, while the DW measures the duration and direction of trends in a country's performance trajectory, capturing early signs of improvement or decline. This approach provides an informative comparison of countries, highlighting not only their current standings but also their trends and trajectories over time. The methodology was used on a combined dataset from various sources to estimate worldwide happiness from 2005 to 2023.

C0877: **Combining quantitative and qualitative assessments in the multidimensional well-being measurement**
Presenter: **Jose Luis Garcia-Lapresta**, Universidad de Valladolid, Spain

A proposal on how to combine quantitative and qualitative assessments in the multidimensional well-being measurement is provided. To this end, some normalization procedures are established in such a way that every assessment is translated into a unit interval. These normalizations allow aggregating assessments issued on various scales, both quantitative and qualitative, in the different dimensions in which the well-being of the citizens of a society is evaluated. In the case of ordered qualitative scales, subjective perceptions of the psychological closeness between the terms of the scales are considered using ordinal proximity measures, as well as functions that associate a numerical value to the terms of the scales based on the mentioned subjective perceptions. Once all the assessments have been normalized into the unit interval, it is possible to aggregate them following "the row-first two-stage aggregation" procedure: first aggregating across dimensions for each individual and then aggregating across individuals. Considering individual's overall achievement scores, it is possible to obtain a social well-being index.

C1076: **Inner areas in Lombardy: An analysis of vulnerability dynamics**
Presenter: **Annamaria Bianchi**, University of Bergamo, Italy
Co-authors: Chiara Gigliarano, Sara Maiorino

The national strategy for "Inner Areas" (SNAI) in Italy represents a pioneering approach to foster development and territorial unity aimed at mitigating marginalization and population decline in these regions nationwide. SNAI is founded on a bold localized strategy, emphasizing new multi-level local governance structures for comprehensive local promotion and development. It seeks to tackle demographic hurdles and cater to the requirements of territories facing significant geographical and/or demographic challenges. In Lombardy, there are 14 inner areas, comprising 488 municipalities and a population of approximately 1.169.000 people. The aim is to analyze some of the fragilities faced by these territories, in terms of services accessibility and economic conditions of its inhabitants. This was done thanks to a set of original data derived from a survey carried out on the topic of the economic conditions of families in Lombardy and two other Italian regions (Tuscany and Campania).

CO196 Room BH (S) 2.03 ADVANCES IN BAYESIAN MACRO- AND FINANCIAL ECONOMETRICS Chair: Toshiaki Watanabe

C0244: **Estimating trend inflation in a regime-switching Phillips curve**
Presenter: **Jouchi Nakajima**, Hitotsubashi University, Japan

A regime-switching Phillips curve model is developed to estimate trend inflation. Extending the earlier work, trend inflation, the slope of the Phillips curve, and the oil price pass-through rate are allowed to follow a regime-switching process. An empirical analysis using Japan's consumer price index illustrates that including the oil price and its time-varying pass-through rate improves the model's ability to forecast inflation. The empirical results also show that the obtained trend inflation highly correlates with firms' inflation expectations.

C0221: **High-dimensional multivariate realized stochastic volatility model using characteristic factor regression**
Presenter: **Tsunehiro Ishihara**, Takasaki City University of Economics, Japan

In financial econometrics, multivariate return series modeling has been widely studied. Multivariate stochastic volatility models are popular and exhibit high forecasting performance. However, estimating and forecasting with multivariate stochastic volatility models is time-consuming, especially for high dimensions. A multivariate stochastic volatility model with observed characteristic factors is proposed. Intraday information is also introduced via realized covariance. In the proposed model, high-dimensional multivariate returns are conditionally independently modeled by a univariate time-varying coefficient regression model with stochastic volatility errors. Estimation and forecasting can be performed separately for each series. Thus, computation time increases proportionally to the dimension of the returns. In addition, estimation and forecasting can be performed in parallel in each univariate model. An empirical illustrative example is presented using sector index series and market-, size-, and value-based factors, as well as their realized covariances.

C0184: **Stochastic volatility in mean: Efficient analysis by a generalized mixture sampler**
Presenter: **Daichi Hiraki**, University of Tokyo, Japan
Co-authors: Siddhartha Chib, Yasuhiro Omori

The simulation-based Bayesian analysis of stochastic volatility in mean (SVM) models is considered. Extending the highly efficient Markov chain Monte Carlo mixture sampler for the SV model proposed in existing studies, an accurate approximation of the non-central chi-squared distribution is developed as a mixture of thirty normal distributions. Under this mixture representation, the parameters and latent volatilities are sampled in one block. A correction of the small approximation error is also detailed by using additional Metropolis-Hastings steps. The proposed method is extended to the SVM model with leverage. The methodology and models are applied to excess holding yields in empirical studies, and the SVM model with leverage is shown to outperform competing volatility models based on marginal likelihoods.

C0644: **Linkage between wage and price inflation in Japan**
Presenter: **Yoichi Ueno**, Bank of Japan, Japan

Changes in the linkage between wages and prices in Japan are investigated using a dynamic factor model of disaggregated wages and prices with heteroscedasticity- and autocorrelation-robust inference. The empirical results show that the model is better at identifying the underlying trends in wage and price inflation than models using only aggregate data. In addition, the trend component of service price inflation is the best indicator to gauge the underlying trend in price inflation among the indicators examined. Further, wages and prices decoupled around 1998, but they have recoupled to some extent in the post-COVID-19 era. Lastly, the volatility of the common trend component of wage and price inflation determines the strength of the linkage between wages and prices, and it closely tracks an indicator which shows the importance of price inflation when firms revise wages in negotiations.

C0722: **Multi-view dynamic network modeling**
Presenter: **Mike So**, The Hong Kong University of Science and Technology, Hong Kong
Co-authors: Shun Hin Chan, Amanda Chu

A flexible multi-view dynamic network model is developed using a regression-like structure, incorporating exogenous and endogenous variables from the lagged networks to model edge changes. The model does not rely on latent space, simplifying network estimation and prediction. Furthermore, it integrates a multi-view feature to represent various relationship types at each time point. The proposed model offers an intuitive interpretation of the estimation. Bayesian model averaging method is also applied to predict networks.

CO307 Room BH (SE) 2.05 FORECASTING: THEORY AND PRACTICE Chair: Daniele Girolimetto

C0454: **Impact of grid innovations on electricity price volatility in Italian island markets**
Presenter: **Pierdomenico Duttilo**, University of Padova, Italy
Co-authors: Francesco Lisi, Marina Bertolini

Electricity market outcomes, specifically prices and volumes, are significantly influenced by infrastructures. In Italy, electricity market zones have unique features due to interconnection levels, energy mixes, and consumption patterns, especially in Sicily and Sardinia, with their limited mainland connections. The 2016 completion of the Sorgente-Rizziconi cable enhanced the link between Sicily and the mainland, affecting day-ahead prices. The impact of grid innovations on the volatility of zonal electricity prices in Sicily is examined by analyzing day-ahead prices and their structural components. In a wide sense, it also forecasts the effects of future grid developments on the Sardinian and Sicilian markets. Studying these island markets is crucial to understanding how infrastructure affects price volatility. These findings can potentially be generalized to other markets with suitable adaptations.

C0464: **Fast forecast reconciliation using sub-hierarchies**
Presenter: **Fotios Petropoulos**, University of Bath, United Kingdom

Hierarchical forecasting has been a prominent research topic for the last 15 years, mainly due to its practical relevance. Various reconciliation methods have been proposed, and these methods offer coherent point and probabilistic forecasts across the various aggregation levels coupled with improved forecasting performance across the hierarchy. However, one main issue of such reconciliation methods is their limited applicability on large hierarchies due to the computational complexity related to the matrix calculations involved. This issue is addressed by proposing an overarching approach to forecast reconciliation methods that is based on the construction of sub-hierarchies. The approach can be applied in conjunction with any known forecast reconciliation method, either for a point or for probabilistic forecasts. Sub-hierarchical forecasting not only renders the reconciliation calculations possible for hierarchies of any size but also results in robust improvements over existing reconciliation methods. The proposed approach is applied to a large hierarchical dataset in the retail context, and its value is showcased in practice.

C0759: **Scalable dynamic hierarchical forecast reconciliation**
Presenter: **Ross Hollyman**, University of Bath, United Kingdom

A dynamic approach is introduced to probabilistic hierarchical forecasting at scale. The model differs from the existing literature in this area in several important ways. Firstly, the weights allocated to the base forecasts in forming the combined, reconciled forecasts are explicitly allowed to vary over time. Secondly, the assumption is dropped, nearly ubiquitous in the literature, that in-sample base forecasts are appropriate for determining these weights and use out-of-sample forecasts instead. Most existing probabilistic reconciliation approaches rely on time-consuming sampling-based techniques and, therefore, do not scale well (or at all) to large data sets. This problem is addressed in two main ways: firstly, by developing a closed estimator of covariance structure appropriate to hierarchical forecasting problems, and secondly, by decomposing large hierarchies into components that can be reconciled separately.

C0902: **On the difference of the existing hierarchical forecasting approaches**
Presenter: **Nikolaos Kourentzes**, University of Skovde, Sweden
Co-authors: George Athanasopoulos

Hierarchical forecasting refers to a class of forecasting problems where different time series are interconnected through aggregation constraints. For instance, cross-sectional hierarchies appear in supply chains, where sales at the store/product level aggregate to sales across various demarcations, such as product categories. Nowadays, the standard approach is to produce independent forecasts for each separate series and reconcile them across the hierarchy to ensure that aggregation constraints are met. The literature is rich with different reconciliation estimators, which, at their core, combine the base forecasts into the reconciled ones. Their quality is typically considered in terms of forecast accuracy. Leveraging the geometric interpretation of forecast reconciliation, it is demonstrated that existing solutions explore a very limited range of the possible solution space. It is first characterized by that space, and then it investigates how these unexplored solutions can be obtained, providing a new class of estimators. These solutions are benchmarked in terms of efficiency and forecasting performance.

C0406: **Coherent forecast combination for linearly constrained multiple time series**
Presenter: **Daniele Girolimetto**, University of Padova, Italy
Co-authors: Tommaso Di Fonzo

When different, incoherent forecasts of a linearly constrained (i.e. hierarchical) multiple time series are available, both forecast combination and forecast reconciliation may be used to improve the forecast accuracy and achieve coherence in the final forecasts. A regression-based optimal solution is presented to the coherent forecast combination problem for $p > 1$ base (i.e., incoherent) forecast vectors of the same target forecast for a linearly constrained multiple time series. Then, practical issues related to the estimation of the covariance matrix on which the optimal solution is based are discussed. Finally, the effectiveness of the proposed approaches is assessed through two forecasting experiments, using the Australian Energy Market Operator (AEMO) and the Australian Tourism Demand datasets.

CO247 **Room BH (SE) 2.09** INEQUALITY AND MACROECONOMIC DYNAMICS **Chair: Nao Sudo**

C1389: **The effects of monetary policy shocks on inequality in Japan**
Presenter: **Tomoaki Yamada**, Meiji University, Japan
Co-authors: Nao Sudo, Masayuki Inui

The impacts of monetary easing on inequality have been attracting increasing attention recently. The micro-level data is used on Japanese households to study the distributional effects of monetary policy. A quarterly series of income and consumption inequality measures are constructed from 1981 to 2008, and their response is estimated to monetary policy shock. It is found that monetary policy shocks do not have a statistically significant impact on inequality across Japanese households in a stable manner. When considering inequality across households whose head is employed, evidence is found that, before the 2000s, an expansionary monetary policy shock increased income inequality through a rise in earnings inequality. Such procyclical responses are, however, scarcely observed when the current data are included in the sample period or when earnings inequality across all households is considered. It is also found that the transmission of income inequality to consumption inequality is minor, including when the procyclicality of income inequality was pronounced. Using a two-sector dynamic general equilibrium model with attached labor inputs shows that labor market flexibility is central to the dynamics of income inequality after monetary policy shocks. The micro-level data on household balance sheets are also used, and distributions of households' financial assets and liabilities are shown to not play a significant role in the distributional effects of monetary policy.

C1404: **The impact of the cost-of-living crisis on European households**
Presenter: **Boris Chafwehe**, Bank of England, United Kingdom
Co-authors: Mattia Ricci, Daniel Stoehlker

The impact of the recent cost-of-living crisis on European households is studied using data on individual consumption, income, and wealth. The various channels are accounted for through which inflation affects individual households and for the monetary and fiscal policy responses to the inflationary shock. Results indicate that, on average, pension-age households lost nearly three times as much as their working-age counterpart due to the devaluation of their nominal wealth. Along the income distribution, differences in nominal asset holdings and in the evolution of nominal incomes imply that the inflationary shock was regressive for working-age households and mostly flat for pension-age households. Overall, high-income working-age households with mortgage debt gained the most from the inflationary surge, while older individuals with large nominal asset positions were those for which the largest losses were recorded. Fiscal policy measures were able to partially offset the impact of the crisis on the most vulnerable households. The interest rate response to the crisis partially offset the losses recorded by households with large nominal asset positions.

C1410: **New golden rule: K and L returns**
Presenter: **Tim Lee**, Queen Mary University of London, United Kingdom

The difference between the risk-free rate (r) and the growth rate (g) of an economy is one of the most important indicators in macroeconomics. A heterogeneous agent is developed, overlapping generations model in which both r and g are endogenous to each other. The growth rate is determined by parents' investment in children's human capital, while the equilibrium interest rate is determined by the level of parents' savings. It is shown that whenever parents invest enough in their children for endogenous growth, the economy is dynamically inefficient. Moreover, longer life spans reduce both r and g while also raising earnings inequality, qualitatively replicating patterns observed in advanced economies in the past half-century. When overlapping generations of workers coexist with capitalists, who accumulate wealth according to a standard model of capital returns heterogeneity and financial frictions, an increase in investment risk or capitalist profits further lowers r, leading to an increase in wealth concentration and "stronger" dynamic inefficiency. Against this backdrop of dynamic inefficiency, the question of whether taxing labor to subsidize capital is an effective tool for fast development in the context of developing (rather than advanced) economies is also revisited.

C1413: **Wealth inequality and economic volatilities**
Presenter: **Hiro Ito**, Portland State University, United States
Co-authors: Joshua Aizenman

The aim is to examine whether and how wealth structure affects the extent of macro stabilization. In countries with higher levels of wealth-income inequality, those with wealth can afford to use their wealth to live in this kind of environment. In contrast, for poorer countries, the costs of public goods can be too expensive. The use of fiscal space may be affected by the degree of wealth distribution. The correlation between the degree of macroeconomic stabilization and the extent of wealth inequality is examined. It is first graphically shown that the sample economies with higher levels of wealth inequality tend to experience higher degrees of inflation volatility, which we do not find in output volatility. Panel regressions also yield findings consistent with graphical examinations. When interaction terms are included between wealth inequality and other macroeconomic variables, it is found that the volatility-increasing impact of wealth inequality is greater for a developing country loaded with a higher level of government debt. The impact of wealth inequality on output growth volatility is lower for commodity exporters but higher for manufacturing exporters. A developing country with more open financial markets tends to experience greater output growth volatility if it has a higher level of wealth inequality.

C1687: **Cyclicality of income growth distribution and the role of monetary policy**
Presenter: **Yaz Terajima**, Bank of Canada, Canada
Co-authors: Carolyn Wilkins

The focus is on the implications of monetary policy for income inequality by asking the following questions: How does monetary policy impact the distribution of household income growth, and how would the policy interact with the business-cycle implications for the distribution? We use comprehensive income information from the Canadian administrative tax records to document how income changes with the business cycle, estimate the impacts of monetary policy shocks on income changes by the income distribution and by the major source of income, and assess how much monetary policy accounts for the income changes across the income distribution. Our findings confirm and broaden those of the literature in that the mean and the skewness of the income-growth distribution are significantly pro-cyclical, while the variance is weakly so and, hence, not counter-cyclical. We find that monetary policy tightening persistently lowers the growth of income for high-income households more than for lower-income households, reducing income inequality. Finally, monetary policy impacts the income of households with non-professional business and investment income as the major income source more than those with labour earnings.

CO233 Room BH (SE) 2.10 RECENT DEVELOPMENTS IN THE ECONOMETRICS OF COMMODITY MARKETS Chair: Malvina Marchese

C0301: **Navigating the French stock market using nonlinear quantitative investing methods**
Presenter: **Julien Chevallier**, IPAG Business School, France
Co-authors: Dinh-Tri Vo

The purpose is to delve into the CAC40, SBF120, and potentially all the stocks listed in the CAC All-Share on Euronext Paris. A nonlinear solver is utilized for portfolio optimization to formulate diverse asset allocation scenarios. This analysis is further strengthened by implementing cutting-edge clustering techniques from the field of machine learning.

C0327: **A novel hybrid ensemble approach to forecast freight rates volatility**
Presenter: **Morten Risstad**, Norwegian University of Science and Technology, Norway
Co-authors: Malvina Marchese, amir alizadeh

The contribution is to the forecasting literature in three important ways. First, the performance of a variety of machine learning algorithms in forecasting freight rates volatility is investigated. Second, an extensive forecasting comparison between traditional GARCH models and machine learning methods is conducted. The aim is not only to make a complete comparison across traditional and ML methods but also to provide evidence of when and why some of these methods improve the accuracy of forecasting volatility. Findings suggest that substantial incremental information about future volatility can be extracted with ML from additional volatility predictors with minimal noise fitting if regularization is applied. In contrast, the GARCH-MIDAS and GARCH-X models yield only minor improvements. However, traditional GARCH models do a better job of capturing the long-range persistence of the volatility. When deep learning models are compared to the benchmark FIGARCH model, the average MSE for FIGARCH is lower, and the crucial driver for this superior performance is the more effective way to capture fractional integration. Thus, the findings prompt the proposal of a novel hybrid ensemble stacking algorithm that combines GARCH models and tree-based algorithms. Its superior forecasting performance is established at several horizons, including 1, 5, 22 and 60 days ahead according to statistical (MCS and SPA tests) and economic loss functions (value at risk).

C0335: **Efficient semiparametric estimation of environmental and climate policy**
Presenter: **Massimiliano Mazzanti**, University of Ferrara, Italy

The aim is to assess the impact of the price of carbon, which is linked to the European market of allowances, on carbon dioxide emissions. To do so, an econometric model is proposed that extends the environmental Kuznets curve (EKC) model in several directions. First, the price of carbon, which is the policy variable, is introduced in the model in a nonparametric fashion; Second, it is proposed to use interactive fixed effects approach to control for latent heterogeneities in both dimensions of panel data; Third, to allow for spatial dependence, spatially correlated errors are introduced. The extended EKC model poses various challenges for estimation. To cope with them, using a profile likelihood approach, a feasible generalized least squares estimator of the parameters of interest is proposed. Furthermore, the policy effects curve is also efficiently estimated. The asymptotic properties of the estimators are shown, and based on these outcomes, the policy effects are empirically evaluated. The approach yields significantly different and more meaningful results compared to those obtained using standard estimation techniques.

C0343: **Valuation of the leasehold properties in the England using a machine learning approach**
Presenter: **Tatiana Franus**, Bayes Business School, City, University of London, United Kingdom
Co-authors: Mark Andrew, James Culley

In England and Wales, the leasehold system divides the ownership of a dwelling into two distinct components for a limited period: leasehold and freehold interests. The leasehold interest pertains to the legal rights of occupation, while the freehold interest encompasses the legal ownership of the land and building. As time passes, the value of the leasehold ownership deteriorates. A novel data-driven approach is introduced to estimate the objective price of the leasehold based on the remaining lease length using a machine learning methodology. Utilizing a dataset of transacted prices from the Land Registry for the period 2010-2016, 183 variables are developed to train machine learning models. The price discounts are estimated for different lease lengths to assess whether they conform to the theoretical prediction that these discounts become steeper as the lease length shortens. From these price discount estimations, the relativity index is then derived, which is the ratio of a finite leasehold value relative to its freehold vacant possession value. The results have the potential to contribute to more equitable outcomes in leasehold extensions, as they enable relativities to be derived from market evidence, providing a more transparent approach to calculating the premium and possessing wide industry implications.

C1187: **Shortages to forecast aggregate and sectoral U.S. stock market realized variance**
Presenter: **Matteo Bonato**, IPAG Business School & University of Johannesburg, Switzerland

Recent global economic and political events have made clear that shortages are a key factor driving macroeconomic and financial market developments. Against this backdrop, the forecasting value of shortages for U.S. stock market realized variance (RV) is studied at the aggregate and sectoral level using data spanning the period 1885-2024 (market) and 1926-2023 (most sectors). To this end, linear and nonlinear statistical learning estimators are considered. When linear estimators (OLS and shrinkage estimators) are used, there is no evidence that aggregate and disaggregate shortage indexes have predictive value for subsequent market or sectoral RVs. In contrast, when random forests are used, a nonlinear nonparametric estimator, aggregate and disaggregate shortage indexes are detected to improve forecast accuracy of market and sectoral RVs after controlling for realized moments (realized leverage, realized skewness, realized kurtosis, realized tail risks). RV is then decomposed into a high, medium, and low-frequency component, and the shortage indexes are found to correlate mainly with the medium and low frequencies of RV.

CO218 Room BH (SE) 2.12 ADVANCES IN PANEL DATA AND CAUSAL INFERENCE Chair: Laura Liu

C0397: **Robust estimation and inference in panels with interactive fixed effects**
Presenter: **Andrei Zeleneev**, University College London, United Kingdom
Co-authors: Timothy Armstrong, Martin Weidner

Estimation and inference are considered for a regression coefficient in panels with interactive fixed effects (i.e., with a factor structure). It is shown that previously developed estimators and confidence intervals (CIs) might be heavily biased and size distorted when some of the factors are weak. Estimators are proposed with improved rates of convergence and bias-aware CIs that are uniformly valid regardless of whether the factors are strong or not. The approach applies the theory of minimax linear estimation to form a debiased estimate using a nuclear norm bound on the error of an initial estimate of the interactive fixed effects. The obtained estimate is used to construct a bias-aware CI, taking into account the remaining bias due to weak factors. In Monte Carlo experiments, a substantial improvement is found over conventional approaches when factors are weak, with little cost-to-estimation error when factors are strong.

C0648: **Time-varying heterogeneous treatment effects in event studies**
Presenter: **Laura Liu**, Indiana University Bloomington, United States
Co-authors: Irene Botosaru

The identification and estimation of heterogeneous treatment effects are studied in event studies. The importance of both heterogeneity in treatment effects and the inclusion of lagged dependent variables are highlighted. Omitting lagged dependent variables can lead to omitted variable bias in the estimation of time-varying treatment effects. Under the assumption of strict exogeneity in the treatment, an empirical Bayes estimator is proposed for the heterogeneous treatment effects, which is flexible and easy to implement. The method also helps shed light on common assumptions in the event study literature, such as the potential correlation between heterogeneous treatment effects and individual heterogeneity, as well as the potential presence of state dependence.

C0719: **To link or not to link: Estimating long-run treatment effects from historical data**
Presenter: **Francis DiTraglia**, University of Oxford, United Kingdom
Co-authors: Ezra Karger, Camilo Garcia Jimeno

A fundamental challenge in empirical research is the reliance on datasets linked with error. Researchers must often match a dataset containing treatment status to a separate dataset with outcome measures, typically relying on non-unique information such as names and demographic characteristics. This imperfect linking process raises serious concerns about statistical efficiency, measurement error, and sample selection. For instance, because nearly all married women in the 1900s changed their surname upon marriage, most research in economic history using linked data excludes women entirely, leaving many important historical questions about them unanswered. A unified method is developed for precisely estimating long-run treatment effects using information from two datasets without explicitly constructing a linked dataset or discarding observations. The approach uses available linking covariates as efficiently as possible, allowing for measurement error in these variables and heterogeneous treatment effects. The method nests the typical approach of using unique matches but extends it to cases where such matches are not universally available. To demonstrate the practical implications of the methodology, it is used to revisit research on compulsory schooling laws and the inter-generational effects of slave-holding on the wealth of slave-owning families.

C1019: **Using multiple outcomes to improve the synthetic control method**
Presenter: **Liyang Sun**, UCL and CEMFI, Spain
Co-authors: Eli Ben-Michael, Avi Feller

When there are multiple outcome series of interest, synthetic control analyses typically proceed by estimating separate weights for each outcome. Estimating a common set of weights across outcomes is proposed instead by balancing either a vector of all outcomes or an index or average of them. Under a low-rank factor model, it is shown that these approaches lead to lower bias bounds than separate weights and that averaging leads to further gains when the number of outcomes grows. This is illustrated via simulation and in a re-analysis of the impact of the Flint water crisis on educational outcomes.

C1367: **Rank assisted network regression**
Presenter: **Weining Wang**, University of Groningen, Netherlands

The aim is to propose studying the interaction effects of social and spatial networks in the presence of a noisy adjacency matrix. First, evidence is provided that existing network datasets exhibit low-rank, sparse, and noisy structures, and this information is utilized to create a de-noised version of the network. The least absolute shrinkage and selection operator (LASSO) is employed in conjunction with nuclear norm penalization to simultaneously regularize the sparse and low-rank components. Two procedures are introduced: a two-step estimator, where the adjacency matrix is first de-noised before using it in regression analysis, and a one-step supervised generalized method of moments (GMM) estimator using proximal gradient methods for efficient computation. Results show that the estimation method performs favorably compared to GMM, especially when dense errors are present and networks are endogenous to measurement errors. Simulation exercises indicate that the method outperforms GMM by up to 30 in root mean squared error (RMSE) terms when noise is present in the network and maintains a significant advantage of approximately 40 on average with endogenous networks.

CC459 **Room S-2.25** **NONPARAMETRIC STATISTICS** **Chair: Enea Bongiorno**

C0926: **Quantile regression with Bernstein polynomials**
Presenter: **Santiago Pereda-Fernandez**, Universidad de Cantabria, Spain

An alternative to quantile regression is proposed to estimate conditional quantiles. To do so, conditional quantiles are modeled using Bernstein polynomials, which are a nonparametric smoother related to series estimation. With this model, it is possible to write the conditional density function in terms of the Bernstein coefficients and the data, which allows to use maximum likelihood for the estimation. Moreover, the estimator has several desirable features, including no quantile crossings, no need to interpolate between quantiles in a grid, the estimated coefficients are differentiable, and simple functions of the coefficients are integrable with respect to the quantile index.

C1315: **Moment-generating function of the Hettmansperger-Norton-type test for patterned alternatives**
Presenter: **Hikaru Yamaguchi**, Tokyo University of Science, Japan
Co-authors: Hidetoshi Murakami

Testing the equality of several location parameters against patterned alternatives, such as ordered or umbrella alternatives, is one of the intriguing subjects in nonparametric statistics. For example, in dose-response studies, the drug effect on animals is expected to increase or decrease as the dosage increases. In such cases, an ordered alternative model is suitable. The Hettmansperger-Norton-type test, defined as a linear combination of linear rank statistics, is a frequently used nonparametric rank test for patterned alternatives. The moment-generating function of the Hettmansperger-Norton-type test is derived in the presence of ties. Furthermore, the explicit formulas of some higher-order moments are also derived to obtain the Edgeworth approximation for the null distribution of the test statistic. Simulation studies demonstrate the usefulness of the Edgeworth approximation by examining the Type I error rate in the case of small sample sizes.

C1441: **Tests for comparing ROC curves under the presence of covariates**
Presenter: **Juan-Carlos Pardo-Fernandez**, Universidade de Vigo, Spain
Co-authors: Aris Fanjul Hevia, Wenceslao Gonzalez-Manteiga

The receiver operating characteristic (ROC) curve is a graphical tool routinely used to evaluate the performance of a binary classification procedure based on a continuous marker. In many practical applications, covariates related to the marker are available. Under these circumstances, it is of interest to evaluate the influence that those covariates might have on the performance of the marker in terms of classification ability by means of the covariate-specific ROC curve, which is defined in terms of conditional distributions. Several tests to compare covariate-specific ROC curves are discussed, including the cases with univariate and multivariate covariates. In practice, these tests would allow to decide if, for a given value of the covariate, the classification capabilities of several markers differ. The proposed methodologies rely on nonparametric estimation of the involved ROC curves and bootstrap resampling plans to approximate the null distribution of the test statistics. The proposed procedures are used to analyze a real data set of patients with pleural effusion.

C1602: **Resampling-based approaches for nonparametric MANOVA in the presence of missing data**
Presenter: **Lubna Amro**, TU Dortmund University, Germany
Co-authors: Markus Pauly

Multivariate analysis of variance (MANOVA) is widely used across various fields to examine multivariate endpoints. Traditional MANOVA methods require complete data and assume multivariate normality and homogeneous covariance matrices, but these assumptions often do not hold. Missing data can complicate these issues, potentially leading to inflated type-I error rates or low statistical power. To address this, resampling-based methods are introduced that handle missing data without the need for imputation or the exclusion of observations. The approach, which integrates resampling with quadratic form-type test statistics, is asymptotically valid and, accommodates heteroscedastic designs and allows for singular covariance matrices. Extensive simulations demonstrate that our methods effectively control type-I error rates and perform well across various distributional scenarios under missing completely at random (MCAR) and missing at random (MAR) mechanisms. Additionally, the methods are applied to a real data example to illustrate their practical applicability.

C1424: **Fast non-parametric test on the equivalence of multivariate empirical distributions**
Presenter: **Johannes Bleher**, University of Hohenheim, Germany

A novel approach is presented for testing the equivalence of two multivariate samples using empirical characteristic functions. Building upon an existing framework in the literature, a test statistic is developed that leverages the unique properties of characteristic functions to detect differences in distribution between two multivariate datasets. The method extends an existing approach, which focuses on goodness-of-fit tests for fully specified theoretical distributions. The approach offers a powerful and flexible tool for various applications in statistics and data analysis. The asymptotic distribution of the proposed test statistic is derived under the null hypothesis, and its power is investigated against a range of alternative hypotheses. Through extensive Monte Carlo simulations, the validity of the test performance is investigated. Especially in scenarios involving high-dimensional data or non-normal distributions, the test statistic may offer a computationally efficient way to test for different distributions. Practical guidelines are also provided for implementing the test. Finally, the approach is illustrated through real-world case studies in finance and biostatistics, showcasing its potential for detecting distributional discrepancies in multivariate datasets.

C1730: **General frameworks for conditional two-Sample testing**
Presenter: **Seongchan Lee**, Yonsei University, Korea, South
Co-authors: Ilmun Kim, Suman Cha

The problem of conditional two-sample testing is studied, which aims to determine whether two populations have the same distribution after accounting for confounding factors. This problem commonly arises in various applications, such as domain adaptation and algorithmic fairness, where comparing two groups is essential while controlling for confounding variables. We begin by establishing a hardness result for conditional two-sample testing, demonstrating that no valid test can have significant power against any single alternative without proper assumptions. We then introduce two general frameworks that implicitly or explicitly target specific classes of distributions for their validity and power. Our first framework allows us to convert any conditional independence test into a conditional two-sample test in a black-box manner, while preserving the asymptotic properties of the original conditional independence test. The second framework transforms the problem into comparing marginal distributions with estimated density ratios, which allows us to leverage existing methods for marginal two-sample testing. We demonstrate this idea in a concrete manner with classification and kernel-based methods. Finally, simulation studies are conducted to illustrate the proposed frameworks in finite-sample scenarios.

CC492 Room S-1.06 CLUSTERING Chair: Ioanna Papatsouma

C1234: **Advancements in finite mixture models and flexible model-based clustering techniques**
Presenter: **Samyajoy Pal**, LMU Munich, Germany
Co-authors: Christian Heumann

The aim is to explore the advancements of modelling multivariate data with finite mixture models and model-based clustering by enhancing flexibility, parameter estimation, and performance across diverse data types. A Dirichlet mixture model (DMM)-based clustering method that utilizes a modified hard EM algorithm and a soft EM algorithm is introduced. The approach outperforms popular clustering algorithms, as demonstrated on both simulated and real-world datasets. Additionally, an alternative parametrization of the Dirichlet distribution is proposed using mean and precision parameters, improving interpretability and estimation accuracy. Innovative estimation techniques, such as Stirling's and moment approximations, provide closed-form solutions that boost model identifiability and computational efficiency, especially in high-dimensional settings. Traditional mixture models are further extended by allowing combinations of identical and non-identical distributions, including mixtures of multivariate skew normal and multivariate generalized hyperbolic distributions. This generalized framework effectively captures complex data structures and accurately identifies underlying patterns. Overall, a robust and flexible toolkit is offered for finite mixture modeling and clustering, advancing the capacity to handle complex data scenarios with improved accuracy and interpretability.

C1542: **A deterministic information bottleneck method for clustering mixed-type data**
Presenter: **Efthymios Costa**, Imperial College London, United Kingdom
Co-authors: Ioanna Papatsouma, Angelos Markos

A plethora of algorithms for cluster analysis have been developed in recent years, with most focusing on just continuous data and not being suitable for mixed-type data sets, that is, consisting of both continuous and categorical variables. The clustering techniques that have been proposed to deal with this heterogeneity treat categorical variables as being of the same type (either nominal or ordinal), and many fail to take into account that certain variables may be completely unrelated to the cluster structure. An information-theoretic approach is presented for clustering mixed-type data based on the deterministic variant of the information Bottleneck algorithm. The proposed method treats different variable types separately and seeks to optimally compress the data into clusters while retaining relevant information about the underlying structure. Furthermore, the selection of hyperparameters associated with this method provides the user with the flexibility of incorporating feature selection within the algorithm. The performance of the approach is compared to that of three well-established clustering methods (KAMILA, K-Prototypes, and partitioning around medoids with Gower's dissimilarity) on simulated and real-world datasets. The results demonstrate that the proposed approach represents a competitive alternative to conventional clustering techniques under specific conditions.

C1580: **Non-negative matrix tri-factorization for multi-view biclustering**
Presenter: **Marina Evangelou**, Imperial College London, United Kingdom
Co-authors: Ella Orme

In a range of application domains, data are collected from different sources (referred to as views) to describe the same objects. This could be the same topics reported on by multiple news outlets or in multiple languages or different types of biological information collected on individuals in a clinical study. These are referred to as multi-view or multi-modal data. Incorporating this multi-view data into models can aid in our ability to extract signals and perform statistical tasks. Multi-view clustering is one such task, seeking to cluster samples using data from multiple sources. The further interest is in identifying the features from each view that are the drivers of the sample clusters. This problem is usually referred to as bi-clustering, where both the rows and the columns of a data matrix are clustered simultaneously. The problem of multi-view biclustering and a novel approach based on matrix factorization, known as restrictive multi-view non-negative matrix tri-factorization (ResNMTF), are discussed. Additionally, an extension of the popular silhouette score allowing for comparisons between biclusterings is discussed.

C1527: **A hybrid approach for the spatial clustering problem via the cross-entropy method**
Presenter: **Nishanthi Raveendran**, Western Sydney University, Australia
Co-authors: Georgy Sofronov

Spatial data is often heterogeneous, meaning a single statistical model may not accurately describe the data. To address this, the data can be divided into several homogeneous regions or domains. The process of identifying these regions and their boundaries is known as spatial clustering (or segmentation) in spatial statistics. Spatial clustering has wide applications across various fields, including criminology, epidemiology, and ecology. The focus is on spatially correlated crime data. Due to the complexity of the model space, which is non-Euclidean after accounting for cluster label permutations, a hybrid approach combining the Cross-Entropy method with a genetic algorithm is proposed. The results demonstrate that the proposed algorithm effectively identifies homogeneous clusters within spatial crime data, providing valuable insights for crime pattern analysis.

CC425 Room S-1.27 MULTIVARIATE METHODS Chair: Alessia Pini

C1282: **The flow copula class**
Presenter: **Maximilian Coblenz**, Ludwigshafen University of Business and Society, Germany
Co-authors: Bolin Liu, Oliver Grothe

A new class of copulas, the flow copula, is introduced based on the change of variables formula and Sklar's theorem. Properties of the flow copula are shown, including the universal expressive power of the class, i.e., that any absolutely continuous copula can be modelled by a flow copula. Furthermore, practical constructions of flow copula models that rely on the normalizing flow technique are presented, and a specific training procedure is proposed that guarantees the copula properties, such as uniform marginal distributions. The proposed method can be used not only to estimate a flow copula model from data without prior knowledge of the copula family but also to generate synthetic data efficiently. It is shown through simulation studies and real data applications that the presented flow copula models can represent various dependence properties such as tail dependence and asymmetry. Moreover, they show results comparable to those of other non-parametric copula estimation methods.

C1357: **Testing the independence of variables for specific covariance structures: A simulation study**
Presenter: **Filipe Marques**, University of Lisbon, Portugal

The purpose is to show how it is possible to test the nullity of covariances in a set of variables using a simple univariate procedure. The methodology proposed enables performing the multivariate test of independence of several variables under specific conditions for the covariance structure. The methodology proposed may be used in the high-dimensional setting and, given its simplicity, allows for overcoming the difficulties in using the exact distribution of the statistic used in the likelihood ratio testing procedure. A simulation study is provided to assess the power and significance level, in different scenarios, of the testing procedure proposed when compared with different likelihood ratio tests and a methodology available in the literature.

C1049: **Multi-attribute preferences: A transfer learning approach**
Presenter: **Sjoerd Hermes**, Wageningen University, Netherlands
Co-authors: Joost van Heerwaarden, Pariya Behrouzi

A novel statistical learning methodology is introduced based on the Bradley-Terry method for pairwise comparisons, where the novelty arises from the method's capacity to estimate the worth of objects for a primary attribute by incorporating data of secondary attributes. These attributes are properties on which individuals evaluate objects in a pairwise fashion. By assuming that the main interest of practitioners lies in the primary attribute and that the secondary attributes only serve to improve the estimation of the parameters underlying the primary attribute, the well-known transfer learning framework is utilised. To wit, the proposed method first estimates a biased worth vector using data pertaining to both the primary attribute and the set of informative secondary attributes, which is followed by a debiasing step based on a penalized likelihood of the primary attribute. When the set of informative secondary attributes is unknown, their estimation is allowed for by a data-driven algorithm. Theoretically, it is shown that, under mild conditions, the ℓ_∞ and ℓ_2 rates are improved compared to fitting a Bradley-Terry model on just the data pertaining to the primary attribute. The favorable (comparative) performance under more general settings is shown by means of a simulation study.

C1517: **Estimation methods of heterogeneous treatment effects extending the w-method and a-learner for multiple outcomes**
Presenter: **Shintaro Yuki**, Doshisha University, Japan
Co-authors: Kensuke Tanioka, Hiroshi Yadohisa

In a two-arm trial, participants are assigned to either the treatment or control group. If the treatment's efficacy is unclear within the overall population, identifying effective subgroups is crucial. This can be done by estimating heterogeneous treatment effects (HTE). While recent advances in HTE estimation use complex models that improve accuracy, they often reduce interpretability. Although methods for single continuous or binary outcomes are well-developed, approaches for multiple outcomes are less common. Current interpretable methods, like the W-method and A-learner, estimate HTE for single outcomes but struggle with the correlation structure in multiple outcomes. The aim is to propose two interpretable HTE estimation methods for multiple continuous and binary outcomes. Multiple squared loss and multiple logistic loss functions are introduced, which are bi-convex and based on reduced-rank regression, capturing correlations among outcomes. The proposed methods extend the W-method and A-learner to handle multiple outcomes. By leveraging the bi-convexity of the loss functions, the methods express HTE by fixing one parameter and optimizing the other. Additionally, it is shown that correcting bias in the traditional A-learner for single binary outcomes enables optimal treatment selection, and subgroup interpretability is enhanced using a group lasso penalty.

C1701: **Response prediction with convergence guarantees on multiple random graphs on unknown manifolds**
Presenter: **Aranyak Acharyya**, Johns Hopkins University, United States
Co-authors: Jesus Arroyo, Michael Clayton, Marta Zlatic, Youngser Park, Carey Priebe

In real life, data in the form of multilayer networks, modeled by graphs on a common set of nodes but with edges forming randomly according to different laws at different graphs, are often encountered. In the real world, high dimensional data often admits an underlying low dimensional manifold structure. In our context, we assume that the multilayer of graphs under consideration is located on an unknown one-dimensional manifold in a higher-dimensional ambient space, with each graph corresponding to a point on the manifold. Under certain regularity assumptions, we propose a method to consistently predict graph level responses at an unlabeled graph, in a semisupervised setting, by exploiting the underlying unknown manifold structure.

CC446 Room K0.18 APPLIED STATISTICS Chair: Joachim Schnurbus

C0823: **Spatial correction of low-cost sensors observations for fusion of air quality measurements**
Presenter: **Jean-Michel Poggi**, University Paris-Saclay Orsay, France
Co-authors: Bruno Portier, Michel Bobbia

The context is the statistical fusion of air quality measurements coming from different monitoring networks. The first one consists of high-quality reference sensors, and the second consists of low-quality micro-sensors. Pollution maps are obtained by correcting the output of numerical models with the measurements from the monitoring stations of the air quality networks. Increasing the density of sensors would then improve the quality of the reconstructed map. Usually, a geostatistical approach is used for the fusion of the measurements, but the first step is to correct the measurements of the micro-sensors, thanks to those given by the reference sensors, by prior offline fitting of a model issued from a costly and sometimes impossible colocation period. The proposal complements these approaches by considering the online spatial correction of microsensors. The basic idea is to use the reference network to correct the measurements of network 2: first, the reference measurements are estimated by kriging only the measurements of network 2; then, the residuals of the estimation on network 1 are calculated; and finally, the correction to be applied to the microsensors is obtained by kriging these residuals. This sequence of steps can then be iterated or not, and the role of the networks can be alternated or not during the iterations. This algorithm is first presented, then studied by simulation and finally applied to a real data set.

C1286: **SWAG: Interpretation, replicability and statistical inference with multiple simple models**
Presenter: **Roberto Molinari**, Auburn University, United States
Co-authors: Stephane Guerrier, Nabil Mili, Yagmur Yavuz-Ozdemir, Samuel Orso, Cesare Miglioli, Gaetan Bakalli

Decision-making based on data analysis goes through an exploratory phase, which usually leads to conclusions that rely on the single "best" interpretation/prediction of the relationships detected in the data. While this approach is standard, in many cases, the uncertainty in the data and/or the presence of latent (unobserved) variables does not justify the reliance on a single model, which may often not respond to the user's needs or may even clash with their domain expertise. Recent literature has focused on the selection of "sets" of simple and accurate models that can be selected according to the needs and expertise of analysts (Rashomon Set Theory). In this direction, a heuristic algorithm is presented called the sparse wrapper algorithm (SWAG), which has been used in various applied studies and addresses different practical needs, starting from how multiple (sparse or simple) models can address issues of replicability and can be interpreted jointly in a network allowing them to be used, individually or jointly, to accurately predict outcomes or optimize different user-defined decision criteria. Newly proposed inference tools for this new multi-model paradigm (Rashomon Inference) are also presented.

C1346: **Benford's law(s) and power law distributions**
Presenter: **Claudio Lupi**, University of Molise, Italy

Benford's law predicts that the relative frequency of the first significant digits in many variables in both the natural and social sciences is not uniformly distributed but rather follows a distribution in which smaller numbers are more likely to occur than larger ones. Although the reasons for the emergence of Benford's law in observed data are still not entirely clear, its empirical validity has been confirmed in many different disciplines. This empirical regularity is often exploited to investigate data quality and the possible presence of data manipulations, e.g. in finance, economics, accounting, and epidemiology. The purpose is to show that Benford's law and power laws are strictly connected. Many Benford random variables display a power law distribution in a delimited range. It should, therefore, come as no surprise that the fields of application of Benford's law and power laws are often very similar. It also shows how to simulate Benford random variables with characteristics (range, quantiles, mean) similar to those observed but not necessarily Benford data. This can prove useful in assessing the Benfordness of the observed data.

C1575: **Estimating treatment decision rules for ordinal outcomes with applications to an antidepressant treatment trial**
Presenter: **Adam Ciarleglio**, George Washington University, United States

In most studies of antidepressant treatment effects, response data are collected at multiple time points, yielding a treatment response trajectory for each subject. In many instances, these trajectories can reasonably be assumed to arise from latent classes that are characterized by both the timing and quality of the response, which can be considered as an ordinal outcome (e.g., non-responders, delayed responders, gradual responders, or early responders). Using data from a randomized controlled trial comparing different antidepressant treatment combinations, we sought to develop and evaluate a treatment decision rule for selecting an optimal combination treatment for falling in a more desirable trajectory class using patient characteristics measured at baseline. A two-stage approach is employed to estimate the decision rule. The first stage fits predictive models in each treatment arm using parametric and flexible non-parametric methods. The second stage uses predictions from the first stage to estimate a generalized odds ratio, which is then modeled as a function of the baseline characteristics, yielding the estimated treatment decision rule. This approach is assessed via a simulation study and applied and evaluated using the antidepressant trial data.

C1305: **Modelling gunfire in Washington, D.C. using a spatiotemporal Hawkes process with nonseparable triggering function**
Presenter: **Tom Stindl**, UNSW, Australia
Co-authors: Jeffrey Kwan, Feng Chen, Yongtao Guan

The United States has an epidemic of gun violence, with the risk of firearm-related death being significantly higher compared to other high-income nations. Implementing intervention strategies to reduce firearm-related activity requires an improved understanding of the temporal, spatial, and contagious dynamics of gunfire. A spatiotemporal Hawkes process is introduced to model the space-time clustering of gunfire in Washington, D.C., in 2018 using data obtained from the ShotSpotter technology. The model has three essential features. First, space-time interactions are introduced for contagious gunfire by modeling the triggering function as a mixture of space-time functions. Second, the productivity of contagious gunfire varies across both time and space. Third, the intensity of non-contagious gunfire incorporates demographic and socioeconomic variables at the census block level to better understand the characteristics that lead to persistent gunfire in particular communities. The intensity also includes cyclic trends such as daily, weekly, and long-term effects. An estimation procedure is proposed within the expectation-maximization framework, and the results are discussed and interpreted.

CC478 Room K0.20 ROBUST STATISTICS Chair: Andreas Artemiou

C0224: **A novel robust weighted least squares regression line**
Presenter: **Alfonso Garcia-Perez**, Universidad Nacional de Educacion a Distancia (UNED), Spain

Some regression line estimators, such as least median squares (LMS) and least trimmed squares (LTS) estimators have been proposed in the literature as robust alternatives to the classical least squares LS estimators. A novel weighted least squares regression line is introduced, based on considering, as weights, the tail distribution of the studentized residuals Tn when errors in the model follow a scale-contaminated normal distribution. Until now, weights in other weighted least squares estimators are based on the squared-residuals magnitude in LS. Tn distribution is used, so the outliers with respect to Tn are less weighted because they are less likely to appear. This new regression line has some advantages, such as smooth and differentiable objective function, contrary to what happens with LMS or LTS. It also has a closed-form expression and good interesting properties. Finally, with this new regression line, a new robust inverse-variance weighted (IVW) estimator in Mendelian randomization is proposed because of the relationship between the IVW estimators and the slopes of the regression of effect on the instrumental variables Z and the regression of Cause on Z.

C1216: **A robust coefficient of determination with the gamma divergence**
Presenter: **Takeshi Kurosawa**, Tokyo University of Science, Japan
Co-authors: Genta Nakane

In linear regressions, many robustification methods have been developed. The M estimation with a phi function is a very effective way to get robustification. A phi function with the redescending property gets robustification against large outliers. The Turkey weight is a very useful function with the redescending property. In linear regressions, an interest in a coefficient parameter estimation problem of explanatory variables is present. In addition to the coefficient parameter estimation, an evaluation of the obtained model is important in the linear models. The coefficient of determination is usually used for the evaluation. There are, however, few studies relating to the robustification of the coefficient of determination. It is insufficient to use the coefficient of determination using the estimate of the coefficient parameter with robustification. The coefficient of determination itself must be introduced with robustification. A few studies exist based on the typical robustification method. However, some methods do not behave well for a large outlier because they have a bias toward the true value of the coefficient of determination. Therefore, a new method is proposed to get robustification using a gamma divergence.

C1281: **Oja depth for object data**
Presenter: **Vida Zamanifarizhandi**, University of Turku, Finland
Co-authors: Joni Virta

The Oja depth (simplicial volume depth) is one of the classical statistical techniques for measuring the central tendency of data in multivariate space. Despite the widespread emergence of object data like images, texts, matrices or graphs, a well-developed and suitable version of Oja depth for object data is lacking. To address this shortcoming, a novel measure of statistical depth is proposed, the metric Oja depth applicable to any object data. Then, several competing strategies are developed for optimizing metric depth functions, i.e., finding the deepest objects with respect to them. Finally, the performance of the metric Oja depth is compared with three other depth functions (half-space, lens, and spatial) in diverse data scenarios.

C1498: **Least trimmed squares: Nuisance parameter free asymptotics**
Presenter: **Bent Nielsen**, University of Oxford, United Kingdom
Co-authors: Vanessa Berenguer Rico

The least trimmed squares (LTS) regression estimator is known to be very robust to the presence of 'outliers'. It is based on a clear and intuitive idea: in a sample of size n, it searches for the h-subsample of observations with the smallest sum of squared residuals. The remaining $n-h$ observations are declared 'outliers'. Fast algorithms for its computation exist. Nevertheless, the existing asymptotic theory for LTS, based on the traditional ε-contamination model, shows that the asymptotic behavior of both regression and scale estimators depends on nuisance parameters. Using a recently proposed new model, in which the LTS estimator is maximum likelihood, it is shown that the asymptotic behavior of both the LTS regression and scale estimators are free of nuisance parameters. Thus, with the new model as a benchmark, standard inference procedures apply while allowing a broad range of contamination.

C1335: **Robust generalized Bayesian inference via divergences for von Mises-Fisher distribution**
Presenter: **Tomoyuki Nakagawa**, Meisei University, Japan
Co-authors: Sho Kazari, Kouji Tahata, Yasuhito Tsuruta

The aim is to consider Bayesian inference based on the robust divergences for von Mises-Fisher distribution. Handling outliers is a common challenge in statistics, as they can significantly influence estimation results, even in directional data. While numerous studies have developed robust estimation methods for directional data with outliers in the frequentist frameworks, there are a few robust estimations in the Bayesian frameworks. The generalized posterior is provided using robust divergence for von Mises-Fisher distribution, and its asymptotic properties and robustness are demonstrated. Furthermore, the posterior computation algorithm is developed by adopting the weighted Bayesian bootstrap.

CC462 Room S0.03 HIGH-DIMENSIONAL STATISTICS AND ECONOMETRICS Chair: Andrew Wood

C0386: **Precision matrix estimation using penalized generalized Sylvester matrix equation**
Presenter: **Vahe Avagyan**, Wageningen University and Research, Netherlands

Estimating a precision matrix is an important problem in several research fields when dealing with large-scale data. Under high-dimensional settings, one of the most popular approaches is optimizing a Lasso or L1 norm penalized objective loss function. This penalization endorses sparsity in the estimated matrix and improves the accuracy under a proper calibration of the penalty parameter. The problem of minimizing Lasso penalized D-trace loss is demonstrated to be solved by solving a penalized Sylvester matrix equation. Motivated by this method, estimating the precision matrix is proposed using penalized generalized Sylvester matrix equations. A particular estimating equation and a new convex loss function constructed through this equation are developed, which is called the generalized D-trace loss. The performance of the proposed method is assessed using detailed numerical analysis, including simulated and real data. Extensive results show the advantage of the proposed method compared to other estimation approaches in the literature.

C1656: **On influential variables driving change points in high dimensional data**
Presenter: **Shrog Albalawi**, Durham university, United Kingdom
Co-authors: Reza Drikvandi

Detection of change points in a sequence of high dimensional observations is a challenging problem, especially when the change is due to a small number of variables, often known as a sparse change point. A question of interest is how to identify the variable or group of variables that caused a sparse change point in high-dimensional datasets. When a change point is detected, we propose a method that identifies the crucial variables driving the change point by grouping variables according to an appropriate distance measure and assessing how those groups contribute to the change point. The approach reduces dimensionality while accurately identifying the most influential variables. By applying a grouping strategy where the number of groups k satisfies $k < n < p$ with n being the number of observations and p the number of variables, the method enhances the interpretability of high dimensional data and provides insights into the factors driving the change points. Through numerical simulations, the performance of the method is illustrated. The results show an improved performance as both p and n increase, even in sparse settings.

C0336: **Interaction screening via Kendall's rank correlation for imbalanced multi-class classification**
Presenter: **Shuntaro Tanaka**, Shiga University, Japan
Co-authors: Hidetoshi Matsui

Screening is a useful method for selecting important variables for high-dimensional data where the number of predictors is much larger than the sample size. Screening can eliminate unnecessary variables at a low computational cost by calculating their importance scores, such as the correlation between the response and predictor variables. The problem of selecting interactions in classification problems for data with imbalanced sample sizes between classes is considered. Specifically, a method is proposed, called class-to-class KIF (CCKIF), to select interactions in imbalanced multi-class classification problems. CCKIF takes the difference in Kendall's rank correlations for each class to calculate the importance scores of the interactions, improving selection accuracy more than the existing method, even for imbalanced data. The theoretical properties of the proposed method are provided. Simulation studies and real data analysis show that the proposed CCKIF appropriately selects important interactions, especially for data on minor classes.

C1382: **Challenges of cross-validation in post-double-Lasso: A Monte Carlo study**
Presenter: **Adrian Drexel**, University of Regensburg, Germany

Monte Carlo simulations evaluate the performance of post-double-Lasso, revealing that using cross-validated λ (CV-λ) in Lasso can be disadvantageous compared to the X-independent λ, particularly in small samples. Additionally, in settings characterized by approximate sparsity, even post-Lasso with CV-λ occasionally outperforms post-double-Lasso with CV-λ. These results highlight the importance of the penalty choice in high-dimensional econometric models.

C1565: **Regularized Wishart autoregressive stochastic volatility**
Presenter: **Roman Liesenfeld**, University of Cologne, Germany
Co-authors: Guilherme Moura

An existing stochastic volatility (WSV) approach to model high-dimensional financial return series is extended with time-varying volatility and correlations. The WSV is a state-space model with a multiplicative evolution of the precision matrix driven by a multivariate beta variate. Predictions of the covariance matrix are weighted moving averages with exponential forgetting that discounts information in past returns uniformly over time and across the space of the return vector. The extension of the WSV regularizes this forgetting scheme by shrinking the covariance matrix predictions toward a pre-specified target matrix. This restricts the variation of the latent covariance matrix, thereby ensuring stationarity of returns and stabilizing eigenvalues of the covariance matrix predictions. Furthermore, this regularized forgetting is combined with time-varying and directional forgetting to achieve robustness to sudden changes in the correlation structure triggered by transitions from one market regime to another. The regularized WSV maintains closed-form sequential updating formulas for filtering, prediction and likelihood evaluation, facilitating practical implementation. To evaluate its performance, a historical dataset of returns of US stocks is used, and the ability is considered to predict the covariance matrix and the weights of the global minimum variance portfolio. The results show that the regularized WSV approach performs well compared to several alternative models.

CC439 **Room S0.12** **EXTREME VALUES** **Chair: Abdelaati Daouia**

C0884: **Asylum seekers at the extremes**
Presenter: **Mohammad Noori**, University of Trento, Italy
Co-authors: Marco Bee

A first-hand analysis is conducted on forecasting asylum-seekers flow across seven major host countries, including the USA, Germany, France, Italy, Spain, Austria, and Greece, using extreme value theory (EVT) methods. Several modeling and forecasting approaches are employed, including the Poisson count regression, generalized Pareto distribution (GPD), Bayesian GPD, and non-stationary models to estimate the expected asylum-seekers flow for each host country by 2027.

C1228: **Efficient composite likelihood estimation for latent spatial max-stable models**
Presenter: **Patrick Osatohanmwen**, Free University of Bozen-Bolzano, Italy

Max-stable processes are a natural extension of multivariate extreme value theory, which is particularly useful for modelling spatial extremes. The modeling of rare binary events observed across a spatial domain through a latent process approach is considered. Inference for latent max-stable processes presents considerable challenges due to the intractability of the full likelihood function. To circumvent this obstacle, an efficient pairwise likelihood estimation procedure which leverages bivariate latent max-stable likelihood functions is developed. To retain statistical efficiency, the proposed methodology includes a truncation procedure designed to alleviate the complexity and computational burden associated with the overall pairwise likelihood function. By dropping a number of noisy pairwise likelihoods, the proposed method achieves a favorable trade-off between computational efficiency and statistical accuracy. The effectiveness of the approach is demonstrated through numerical simulations and an application to real-world data, illustrating its utility in practical settings.

C1313: **Estimation of the extreme value index with probability weighted moments**
Presenter: **Frederico Caeiro**, NOVA.ID.FCT - Universidade Nova de Lisboa, Portugal
Co-authors: Ivette Gomes

In statistics of extremes, the estimation of the extreme value index (EVI) is an important and central topic of research. The probability-weighted moment estimator of the EVI is considered based on the largest observations of a Pareto-type model. Due to the specificity of the properties of the estimator, a direct estimation of the threshold is not straightforward. An adaptive choice of the number of order statistics to be used in the estimation is considered. The introduced methodology is also applied to a real data set.

C1400: **Estimation of the extreme value index using generalized probability weighted moments**
Presenter: **Ayana Mateus**, NOVA.ID.FCT - Universidade Nova de Lisboa, Portugal
Co-authors: Frederico Caeiro

In the field of statistics of extremes, precise estimation of the extreme value index is essential for accurate tail inference. This parameter enables the estimation of other tail-related parameters, such as extreme quantiles, which provide critical information for decision-makers in industries ranging from insurance and finance to environmental management and engineering. Pareto-type models are examined within a semi-parametric framework, introducing a new class of estimators based on a generalized probability-weighted moments. The asymptotic distribution of the proposed class of estimators is established, and illustrations based on simulated values are provided.

C1582: **Statistical analysis and modelling of extremes in time series**
Presenter: **Clara Cordeiro**, FCiencias.ID, Associacao para a Investigacao e Desenvolvimento de Ciencias (Portugal), Portugal
Co-authors: Dora Prata Gomes, Celestino Coelho, Manuela Neves

The statistical analysis of extreme values has gained prominence in studying events such as floods, heatwaves, hurricanes, and sea level rise, among many others. These phenomena have led to the development of special statistical methodologies for their investigation, understanding, and control where feasible. Such events present distinct statistical challenges, requiring a proper characterisation of the tail of the distribution of the variable under study. Extreme value theory (EVT) is the branch of statistics dedicated to modelling these data types. Conventional forecasting methods need to account for these extreme events effectively, as demonstrated by limitations in their extrapolation. A proposed method involves applying a forecasting procedure to the time series and modelling the residuals with a suitable EVT distribution. In addition, the extreme distribution of the residuals is estimated, and bootstrap estimators of the shape parameter are used to enhance the tail behavior of the residual distribution. A simulation study is presented, and a real case study illustrates the contribution to the modelling of extremes in time series analysis.

CC454 **Room S0.13** **BIOSTATISTICS** **Chair: Federico Camerlenghi**

C1259: **Patient risk profiling with pair-copula constructions**
Presenter: **Ozge Sahin**, Delft University of Technology, Netherlands

Patient-risk profiling is crucial for optimizing post-operative care and resource allocation, yet traditional binary metrics often fail to capture the full spectrum of patient risk. Pair-copula constructions (PCCs) are applied to model complex dependencies among mixed continuous-discrete clinical variables. Posterior probabilities of outcomes are estimated through discriminant analysis with PCCs and tailored selection and estimation methods. A data-driven framework is introduced to define patient risk groups based on these probabilities. Using a colorectal and small bowel surgery dataset, the method is evaluated against established clinical benchmarks, demonstrating its effectiveness in identifying low-risk patients and highlighting challenges in predicting high-risk cases. The value of PCCs in enhancing predictive accuracy and supporting clinical decisions is shown, such as safe patient discharge.

C1301: **Subject-level segmentation precision weights for volumetric studies involving label fusion**
Presenter: **Christina Chen**, University of Pennsylvania, United States
Co-authors: Sandhitsu Das, Matthew Tisdall, Fengling Hu, Andrew Chen, Paul Yushkevich, David Wolk, Russell Shinohara

In neuroimaging research, volumetric data contribute valuable information for understanding brain changes during both healthy aging and pathological processes. Extracting these measures from images requires segmenting the regions of interest (ROIs), and many popular methods accomplish this by fusing labels from multiple expert-segmented images called atlases. However, post-segmentation, current practices typically treat each subject's measurement equally without incorporating any information about the variation in their segmentation precision. This naive approach hinders comparing ROI volumes between different samples to identify associations between tissue volume and disease or phenotype. A novel method is proposed that estimates the variance of the measured ROI volume for each subject due to the multi-atlas segmentation procedure. It is demonstrated in real data that weighting by these estimates markedly improves the power to detect a mean difference in hippocampal volume between controls and subjects with mild cognitive impairment or Alzheimer's disease.

C1314: **Network meta-analysis of diagnostic test accuracy reported at multiple thresholds**
Presenter: **Efthymia Derezea**, University of Bristol, United Kingdom
Co-authors: Hayley Jones

Network meta-analysis of diagnostic test accuracy (NMA-DTA) is a relatively new field involving combining evidence across studies to evaluate and compare the accuracy of different tests for a given condition. Many commonly used diagnostic tests are continuous biomarkers whose accuracy is evaluated at multiple thresholds within a study. Using current NMA-DTA methods it is feasible to include in an analysis only a few thresholds per study. An approach that can efficiently encompass all available data is discussed. This is a hierarchical model that incorporates multinomial likelihoods for studies reporting results across multiple thresholds and a parametric structure for the relationship between the probability of testing positive and the threshold within each disease class. This approach enables obtaining the accuracy estimates of tests across the whole range of observed thresholds while retaining all the useful properties of standard NMA-DTA methods. Different variations of this model are explored based on the inclusion of study-level random effects and the addition of a further hierarchical structure on the test-level variance components. This method is applied to data from a systematic review of the accuracy of tests for hepatocellular carcinoma in patients with liver cirrhosis.

C1392: **Transcriptomics from the perspective of spatial statistics: Challenges and methodological approaches**
Presenter: **Jose Luis Romero Bejar**, University of Granada, Spain
Co-authors: Juan Manuel Praena Fernandez, Francisco Javier Esquivel

With the advancement of new bioinformatics technologies, some research has focused on the analysis of the sequencing of genetic material within the framework of omics sciences. The analysis of differential gene expression using multivariate techniques applied to high-dimensional data analysis helps to identify significant differences between samples, highlighting genes that may play key roles in specific biological conditions or diseases. Furthermore, if spatial information about the data was also obtained, spatial statistics could be extremely useful. Spatial statistics involves analyzing spatial data to understand patterns, structures, and relationships within them. Investigating the spatial dependency structure between some specific genetic information in a sample of cells from the same tissue would allow identifying patterns of organization and relationships between cells that could be crucial for understanding complex biological processes, such as tissue development, disease progression, or response to treatments. Understanding how cells interact spatially within their environment can provide insights into how tissues function and how abnormalities, such as cancer and others, develop and spread. The focus is on the recent paradigm of spatial transcriptomics and the potential possibility of bringing together the well-founded field of spatial statistics.

C1393: **Multivariate techniques for high-dimensional analysis of genomic data**
Presenter: **Francisco Javier Esquivel**, University of Granada, Spain
Co-authors: Juan Manuel Praena Fernandez, Jose Luis Romero Bejar

Molecular biology and medicine are living in an era of exponential advances in which the emergence of the so-called omics sciences is the result of the existence of new technology that allows us to see where it was previously impossible. There are as many omics sciences as there are biological or molecular elements that can be studied by these technologies. Genomics is the study of an organism's genome, i.e. its DNA, and how that information is applied. All living things, from unicellular bacteria to multicellular organisms, such as plants and animals, have DNA. The study of genetics helps to understand how genes work and what impact they have on diseases. Therefore, genomics, together with statistical analysis techniques of complex data, are essential to personalized medicine and early diagnosis of diseases. A review of the state of the art related to multivariate techniques commonly used in this context is performed, and different illustrations of application are addressed using different packages of the Bioconductor project.

CC443 Room BH (S) 2.05 ASSET PRICING Chair: Alessandra Amendola

C1351: **Asset pricing of defining carbon emission targets**
Presenter: **Andreas Stephan**, Linnaeus University, Sweden
Co-authors: Maziar Sahamkhadam, Hans Loof, Petter Dahlstrom

The purpose is to provide evidence that public commitments to reduced carbon emission targets are reflected in stock returns. Utilizing a sample of 6,884 international stocks from 57 countries, including 1,065 firms committed to the Science Based Targets initiative (SBTi), it is shown that at the firm level, such commitments increase the annualized returns by 2.38%, while setting carbon reduction targets leads to an increase of 3.16%. These results remain significant even after controlling for firm characteristics related to carbon emissions. While these firm-level carbon target premiums are significant in developed markets, only the commitment premium is significant in emerging markets. Regarding market-based premiums, from January 2018 to July 2024, long-short portfolios based on SBTi commitments and target-setting generate annualized returns of 7% and 8.17%, respectively. The classical Fama-French risk factors are also demonstrated to not explain these systematic premiums. Applying the Fama-MacBeth cross-sectional regressions, results indicate that the price of carbon target risk is negative.

C1442: **3D-PCA in foreign exchange markets**
Presenter: **Moritz Dauber**, University of Innsbruck, Austria

The method of 3D-PCA is applied to a cross-section of currency portfolios to analyze the driving forces of risk premia in currency markets. 3D-PCA is a dimension reduction technique to extract (latent) factors based on a tensor model decomposition. One strong advantage of the 3D-PCA estimation over standard PCA is its robustness, even in short samples, leading to a particularly superior out-of-sample performance of the derived factors. A cross-section of currency portfolios is formed based on univariate sorts of various characteristics, such as the forward discount, the real exchange rate, momentum or external imbalances, and it is found that the latent factors extracted from 3D-PCA outperform standard PCA factors in terms of cross-sectional pricing errors as well as Sharpe ratios, particularly out-of-sample. Besides that, 3D-PCA allows conclusions regarding the economic interpretation of the (latent) factors to be drawn. Namely, the building blocks of the factors exhibit typical level, slope and curvature patterns of the underlying characteristics. A comparison to well-established factors from the literature shows that 3D-PCA factors perform reasonably well and yield overall similar pricing errors and Sharpe ratios.

C1526: **US equity announcement risk premia**
Presenter: **Lukas Petrasek**, Charles University Prague, Czech Republic
Co-authors: Jiri Kukacka

The announcement risk premia is analyzed on the US market. Previous studies have found that a significant portion of the overall risk premia is earned on FOMC meeting days and on days when inflation and employment reports are published. The evidence suggests that while the announcement risk premium for these days still exists, there is a much wider range of macroeconomic data releases to consider. It is found that between September 1987 and March 2023, 99% of the overall cumulative risk premia on the Russell 3000 index is earned on days when data on 17 important macroeconomic variables are released (46% of all trading days). The average return on those days is 6.7 bps compared to 0.9 bps earned on days without any announcements. A trading strategy that holds long positions in equities on announcement days and long positions in risk-free assets on non-announcement days has a more than two times higher Sharpe ratio over a simple buy-and-hold strategy on equities. Up to 28 bps monthly returns are also documented on market-neutral portfolios sorted based on announcement sensitivity. These results are robust to the inclusion of several controls and are both economically and statistically significant.

C1634: **Treating inflation targeting as a natural option formula**
Presenter: **Yedidya Rabinovitz**, University of Warsaw, Poland

Three novel analytical option formulas are developed that measure inflation given a central bank has an inflation-targeting policy. To define the option function, the inflation targeting rate is stated as a natural option, considering this condition is a natural experiment. Instead of measuring the dynamics via the risk-neutral or arbitrage-free market valuation, this option formula proposes the quantity theory of money as the underlying macroeconomic dynamics of prices, given the elements are independent of risk preferences. In the same manner, the quantity theory of money is proven to be a martingale. The methodology has two steps; in the first step, the pricing formula is derived by the multidimensional Its lemma; in the second step, when long memory exists, the multi-fractional multidimensional formulas are presented for a constant H and non-constant h. Hence, a new method of evaluating a natural experiment as an option is formed.

C1679: **Cross-asset value**
Presenter: **Julien Royer**, CREST, France
Co-authors: Florian Ielpo

A novel methodology is introduced for constructing a cross-asset value strategy through a time-series analysis, leveraging the fundamental premise that asset prices provide the most straightforward measure of valuation. We employ a classical trend-cycle decomposition of the logarithm of asset prices, approached in a model-driven manner. Various methodologies are explored for extracting value signals, with a particular focus on Hamilton's trend/cycle decomposition, to determine its efficacy in deriving value measures. These measures are subsequently utilized to construct a cross-asset value factor, applying consistent metrics across different asset classes. Our methodology is detailed within a general framework, accompanied by diverse empirical applications that demonstrate the practical utility of our approach. This strategy addresses the challenges analysts face in comparing valuations across different asset classes by proposing a unified method to compute such a value measure. The proposed approach is versatile, suitable for application across a broad spectrum of assets and risk premia.

CC501 **Room BH (SE) 2.01** **ADVANCES IN ECONOMETRICS AND FINANCIAL MODELLING** **Chair: Masayuki Hirukawa**

C1522: **Random dynamical systems in a statistical arbitrage strategy on the stock market**
Presenter: **Przemyslaw Jasko**, Krakow University of Economics, Poland

A statistical arbitrage is a long-short, market-neutral trading strategy. The first aim of the work is to mathematically establish structures of random dynamical systems and their generators (stochastic difference equations) representing the movement of prices of assets, such that these structures enable the pursuit of a statistical arbitrage strategy based on the modeled dynamics of the prices of related stocks. The second aim is to empirically find multivariate stochastic processes of related stock prices, forming a random dynamical system whose properties allow us to pursue a statistical arbitrage strategy based on it. The first aim is realized based on the theory of random dynamical models. As tools to find related processes of asset prices, the following tests are used: Breitungs tests for linear cointegration, RCC tests for nonlinear cointegration, and rank test for monotonic cointegration. Then, Bayesian model is presented for time-varying cointegration TVP-VECM-SV with shrinkage priors, and SAVS post-sparsification. The dataset includes 21 log-prices time series of WIG20 stock index and its constituents. For selected subsets of assets selected by cointegration tests, TVP-VECM-SV models are built with shrinkage priors and post-sparsification of the cointegration matrix. Such model structure enables the simultaneous test if cointegration is present, and when true, it is time-varying (with possible time subperiods in which cointegration disappears) or time-constant.

C1518: **Minimum wage and inflation in European Union countries**
Presenter: **Aleksandra Majchrowska**, University of Lodz, Poland
Co-authors: Sylwia Roszkowska

The purpose is to investigate the impact of minimum wage increases on inflation rates in European countries. The sample includes all EU countries with a national minimum wage. The research period is from 2003 to 2023, and the statistical data source is the Eurostat database. To verify the transmission effect of the minimum wage on prices, a minimum wage-augmented New-Keynesian Phillips curve is used. Dynamic panel models are employed. The results show that a minimum wage increase translates into higher inflation rates. The effects of increases in the minimum wage vary both over time and between countries, as well as within groups of primary products or services. Increases in the minimum wage are more significant in periods of high inflation. The minimum wage produces greater inflationary pressures in countries with strong labor markets and relatively high wages. Companies in these countries are able to transfer more of the increase in labor costs to consumers. The results are particularly relevant to labor market policy. They reveal that even if an increase in the minimum wage does not involve a decrease in employment, it can generate inflationary pressures.

C1708: **A global look into stock indices for dividend payout, corporate cash, and ESG issues**
Presenter: **Kei-Ichiro Inaba**, Hitotsubashi University Business School ICS, Japan

By conducting country-level panel-data regressions to investigate the determinants of divided payments in 18 countries' representative stock market indices over the period 2008-2020, we analyze the trade-off in the distribution of net profits between paying dividends and retaining the rest in relation to environment (E), social (S), and governance (G) issues. Country-specific E, S, and G factors are formulated to work as the moderator variables for dividend payments. We find that dividend payments at a given year were positively associated with annual net profits, cash holdings at the beginning of the year, and leverage at the end of the year. The positive association with leverage decreased its impact as the dependency of national financial systems on banks increased. The impact of the positive association with the profits was weakened by a composite indicator of country-specific E attainments and an indicator for anti-corruption related to the agency costs of equity. Higher E attainments and less corruption were associated with fewer dividend payments in relation to the profits. These associations were the most impactful on dividend payments amongst the regressors. A G factor of protecting minority shareholders helped them become less impactful, resulting in more dividend payments.

C1710: **Piecewise linear solutions for non-stationary models**
Presenter: **Inna Tsener**, Universitat de les Illes Balears, Spain
Co-authors: Mariano Kulish

The aim is to assess the accuracy and efficiency of piecewise linear solutions for non-stationary models with rational expectations. We compare piecewise linear solutions against accurate global solutions. Using the canonical stochastic growth model we show that the piecewise linear solution is accurate when expansion sequences evolve according to the non-stochastic growth path of the non-linear model and when agents anticipate this growth path.

C1714: **A new approach to constructing probabilistic forecasts with smoothing quantile regression**
Presenter: **Bartosz Uniejewski**, Wroclaw university of Science and Technology, Poland

Accurate short-term price forecasting is essential for daily operations in electricity markets. This article introduces a new method, called Smoothing Quantile Regression (SQR) Averaging, that improves upon well-performing probabilistic forecasting schemes. To demonstrate its utility, a comprehensive study is conducted on two electricity markets, including recent data covering the COVID-19 pandemic and the Russian invasion of Ukraine. The performance of SQR Averaging is evaluated both in terms of reliability and sharpness measures, and economic benefits from a trading strategy. The latter utilizes battery storage and sets limit orders using selected quantiles of the predictive distribution. SQR Averaging leads to profit increases of up to 3.5% on average compared to the benchmark strategy based solely on point forecasts. This is strong evidence for the practical value of using probabilistic forecasts in day-ahead power trading, even in the face of the COVID-19 pandemic and geopolitical disruptions.

Saturday 14.12.2024 13:40 - 15:20 Parallel Session C – CFECMStatistics2024

CO261 Room S-2.25 COPULAS: METHODOLOGY AND APPLICATIONS Chair: Radu Craiu

C0169: **Index-mixed copulas**

Presenter: **Marius Hofert**, The University of Hong Kong, Hong Kong

The class of index-mixed copulas is introduced, and its properties are investigated. Index-mixed copulas are constructed from given base copulas and a random index vector and show a rather remarkable degree of analytical tractability. The analytical form of the copula and, if it exists, its density are derived. As the construction is based on a stochastic representation, sampling algorithms can be given. A particularly interesting feature of index-mixed copulas is that they allow one to provide a revealing interpretation of the well-known family of Eyraud-Farlie-Gumbel-Morgenstern (EFGM) copulas. Through the lens of index-mixing, one can explain why EFGM copulas can only model a limited range of concordance and are tail-independent, for example. Index-mixed copulas do not suffer from such restrictions while remaining analytically tractable.

C0456: **On factor copula-based mixed regression models**

Presenter: **Bouchra Nasri**, University of Montreal, Canada
Co-authors: Pavel Krupskiy, Bruno Remillard

A copula-based method for mixed regression models is proposed, where the conditional distribution of the response variable, given covariates, is modelled by a parametric family of continuous or discrete distributions, and the effect of a common latent variable pertaining to a cluster is modelled with a factor copula. It is shown how to estimate the parameters of the copula and the parameters of the margins and find the asymptotic behavior of the estimation errors. Numerical experiments are performed to assess the precision of the estimators for finite samples. An example of an application is given using COVID-19 vaccination hesitancy from several countries.

C0470: **Approximate Bayesian computation for factor copula models**

Presenter: **Clara Grazian**, University of Sydney, Australia
Co-authors: Feng Chen, Hanwen Xuan

Analyzing and modelling high-dimensional data has attracted great interest in statistics, particularly in applications relevant to time-series analysis and econometrics. In recent years, people have attempted to combine the ideas from the literature on factor analysis and copulas theory together to build up the so-called factor copula models. The use of factor copula models has been growing due to their ability to explain the dependence structure of high dimensional variables in terms of a few latent factors. This feature saves a large amount of computational burden and provides a decent alternative for analyzing high-dimensional datasets. The focus is on the factor copula model proposed in another study, where people could incorporate the class of dynamic factor models proposed in the literature of time series analysis with arbitrary marginal distributions. Their proposed factor copula models are extended into a Bayesian framework by using approximate Bayesian computation (ABC) methods to replace the simulation-based estimation procedures. It enables to not only overcome the issues of lacking closed-form solution in the factor copulas but also capture the model parameter uncertainties and enhance the predictions. The performance of our Bayesian estimation method is examined in both a simulation study and a real time series dataset.

C1051: **Latent variable models with copulas**

Presenter: **Radu Craiu**, University of Toronto, Canada
Co-authors: Robert Zimmerman

Latent variable models are ubiquitous in the statistical modelling of dynamic systems or when the variable of interest is not directly observable, and one must rely on surrogate measurements. A copula-based generalization is presented, in which the joint distribution of a bivariate surrogate measure depends on the latent variable. A Bayesian model is discussed, and a computational algorithm is offered to sample the posterior. The method is illustrated using numerical experiments and data analysis.

CO328 Room S-1.01 HITEC: THEORY AND APPLICATIONS IN FUNCTIONAL STATISTICS Chair: Enea Bongiorno

C0707: **Extremile scalar-on-function regression with application to climate scenarios**

Presenter: **Maria Laura Battagliola**, Instituto Tecnologico Autonomo de Mexico, Mexico
Co-authors: Martin Bladt

Extremiles provide a generalization of quantiles which are not only robust but also have an intrinsic link with extreme value theory. An extremile regression model is introduced tailored for functional covariate spaces. The estimation procedure turns out to be a weighted version of local linear scalar-on-function regression, where now a double kernel approach plays a crucial role. Asymptotic expressions for the bias and variance are established, applicable to both decreasing bandwidth sequences and automatically selected bandwidths. The methodology is then investigated in detail through a simulation study. Furthermore, the applicability of the model is highlighted through the analysis of data sourced from the CH2018 Swiss climate scenarios project, offering insights into its ability to serve as a modern tool to quantify climate behavior.

C0572: **A RKHS-based Bayesian approach to functional regression**

Presenter: **Antonio Coin**, Universidad Autonoma de Madrid, Spain
Co-authors: Jose Berrendero, Antonio Cuevas

A novel Bayesian methodology is proposed for inference in functional linear and logistic regression models based on the theory of reproducing kernel Hilbert spaces (RKHS's). General models are introduced that build upon the RKHS generated by the covariance function of the underlying stochastic process and whose formulation includes, in particular cases, all finite-dimensional models based on linear combinations of marginals of the process, which can collectively be seen as dense subspace made of simple approximations. By imposing a suitable prior distribution on this dense functional space, data-driven inference is performed via standard Bayes methodology, estimating the posterior distribution through reversible jump Markov chain Monte Carlo methods. In this context, the contribution is two-fold. First, a theoretical result is derived that guarantees posterior consistency based on an application of a classic theorem of Doob to the RKHS setting. Second, it is shown that several prediction strategies stemming from the Bayesian procedure are competitive against other usual alternatives in both simulations and real data sets, including a Bayesian-motivated variable selection method.

C0799: **Minimax estimation for FPCA on discretized data**

Presenter: **Nassim Bourarach**, Universita Paris Dauphine - PSL, France
Co-authors: Vincent Rivoirard, Angelina Roche, Franck Picard

The purpose is to consider p noisy evaluations of n realizations of random functions on a common design $(t_j)_{j=1}^{p} \in [0,1]$,$Y_i\left(t_j\right) := X_i\left(t_j\right) + \varepsilon_{i,j}$ for $(i,j) \in [\![1,n]\!] \times [\![1,p]\!]$, where $\varepsilon_{i,j}$ are i.i.d as $\mathcal{N}\left(0,\sigma^2\right)$ with $\sigma^2 > 0$. The $\varepsilon_{i,j}$'s are independent of the random functions X_i which are also i.i.d. and defined on $[0,1]$. The interest is in the estimation of $(\psi_\ell^*, \lambda_\ell^*)$, respectively the ℓ-th eigenfunction and eigenvalue of the covariance integral operator associated to X. The first contributions are non-asymptotic minimax lower-bounds for the estimation of these eigenelements when the covariance kernel is m-Holder regular (for all $m \in \mathbb{R}_+^*$) and when the spectrum of the covariance operator obeys some constraints. The class of processes used for the minimax study allows us to analyze the impact of the spectrum of the covariance operator on the estimation rates and obtain inconsistent results if the constraints are not satisfied. Then, simple estimators of the eigenelements are presented based on a projection onto a wavelet basis.

The obtained estimators are minimax optimal under additional assumptions and attain rates (in n and p) of the form $n^{-1}+p^{-2m}$. Surprisingly enough, even if the problem is non-parametric in nature, there is actually no need for data smoothing.

C1094: **Some properties of ICA in infinite-dimensional settings**
Presenter: **Marc Vidal**, Ghent University, Max Planck Institute for Cognitive and Brain Sciences, Belgium
Co-authors: Ana Maria Aguilera

Independent component analysis (ICA) is discussed in a setting where infinitely many statistically independent components are allowed. A critical aspect of ICA models is the mixing operator, which turns out to be severely unidentified and ill-conditioned in the current framework. We elaborate on the notion of Hilbertian independence and separability to characterize this operator. Furthermore, it is shown how ICA based on kurtosis can be used to classify functions with near-perfect accuracy and explain the underlying principles of this phenomenon, which have a probabilistic interpretation by the Feldman-Hajek dichotomy. The usefulness of the methods is exemplified through neurophysiological data to identify cortical regions involved in depression disorder.

CO140 Room S-1.04 BRANCHING AND RELATED PROCESSES I Chair: Miguel Gonzalez Velasco

C1521: **Markov branching processes with infinite mean immigration**
Presenter: **Maroussia Slavtchova-Bojkova**, Sofia University, Bulgaria
Co-authors: Penka Mayster

Continuous-time branching processes with immigration were first introduced by B. A. Sevastyanov. In these models, immigration occurs according to a homogeneous Poisson process. Specifically, at the jump points of the Poisson process, a random number of new individuals (or particles) arrive, which then reproduce independently according to a continuous-time Markov branching process (CTMBP). The goal is to study the impact of immigration characterized by infinite mean and variance, driven by a discrete-stable compound Poisson process within the framework of CTMBP. In particular, the aim is to derive the explicit form of the probability-generating function for these processes, as well as the probability of extinction. Ultimately, the limiting behavior of such processes is analyzed.

C1431: **Continuous-time Markov chain models with time-dependent rates**
Presenter: **Fatima Palacios Rodriguez**, Universidad de Sevilla, Spain
Co-authors: Antonio Gomez Corral, Martin Lopez Garcia

Many diseases exist that present a certain seasonality in their behavior. To model these infections, infections rates need to be considered, which take into account the time variable. To this end, a continuous-time Markov chain model with time-dependent rates is introduced. Particularly, an erlangization method is applied to the homogeneous quasi birth-death process. The most important characteristics associated with the proposed model are studied. For instance, the hitting times and the hitting probabilities are calculated.

C1669: **Sharp large deviations for branching process with immigration**
Presenter: **Fengnan Deng**, George Mason University, United States
Co-authors: Anand Vidyashankar

The focus is on the study of sharp large deviation estimates for Galton-Watson branching processes allowing immigration. Conditional and unconditional deviation bounds are derived for estimators of the offspring mean and variance and describe extensions to controlled branching processes under appropriate moment conditions. For this reason, local limit theorems are developed under a finite offspring and immigration mean hypothesis. The primary technical tools employed include Fourier analysis of the generating functions.

C1577: **On Bayesian estimation via divergences for controlled branching processes**
Presenter: **Ines M del Puerto**, University of Extremadura, Spain
Co-authors: Miguel Gonzalez Velasco, Anand Vidyashankar, Carmen Minuesa

The focus is on Bayesian inferential methods for the parameters of interest in controlled branching processes that account for model robustness through the use of disparities. The offspring distribution is assumed to belong to a very general one-dimensional parametric family, and the sample given by the entire family tree up to some generation is observed. In this setting, the D-posterior density is defined, a density function which is obtained by replacing the log-likelihood in the Bayes rule with a conveniently scaled disparity measure. The expectation and mode of the D-posterior density, denoted as EDAP and MDAP estimators, respectively, are proposed as Bayes estimators for the offspring parameter, emulating the point estimators under the squared error loss function or under 0-1 loss function, respectively, for the posterior density. Under regularity conditions, the established EDAP and MDAP estimators are consistent and efficient under the posited model. Additionally, it is shown that the estimates are robust to model misspecification and the presence of aberrant outliers. To this end, several fundamental ideas are developed, relating minimum disparity estimators to Bayesian estimators built on the disparity-based posterior, for dependent tree-structured data.

CO239 Room S-1.06 RECENT ADVANCES IN HIGH-DIMENSIONAL STATISTICAL LEARNING Chair: Tianying Wang

C1249: **The non-overlapping statistical approximation to overlapping group lasso**
Presenter: **Tianxi Li**, University of Minnesota, United States

The group lasso penalty is widely used to introduce structured sparsity in statistical learning, characterized by its ability to eliminate predefined groups of parameters automatically. However, when the groups overlap, solving the group lasso problem can be time-consuming in high-dimensional settings due to the group's non-separability. This computational challenge has limited the applicability of the overlapping group lasso penalty in cutting-edge areas, such as gene pathway selection and graphical model estimation. The purpose is to introduce a non-overlapping and separable penalty designed to efficiently approximate the overlapping group lasso penalty. The approximation substantially enhances the computational efficiency in optimization, especially for large-scale and high-dimensional problems. It is shown that the proposed penalty is the tightest separable relaxation of the overlapping group lasso norm within the large family of norms. Moreover, the estimators derived from the proposed norm are statistically equivalent to those derived from the overlapping group lasso penalty in terms of estimation error, support recovery, and minimax rate under the squared loss. The effectiveness of the method is demonstrated through extensive simulation examples and a predictive task of cancer tumors.

C1303: **Beta regression for double-bounded response with correlated high-dimensional covariates**
Presenter: **Jianxuan Liu**, Syracuse University, United States

Continuous responses measured on a standard unit interval (0,1) are ubiquitous in many scientific disciplines. Statistical models built upon a normal error structure do not generally work because they can produce biased estimates or result in predictions outside either bound. In real-life applications, data are often high-dimensional, correlated, and consist of a mixture of various data types. Little literature is available to address the unique data challenge. A semiparametric approach is proposed to analyze the association between a double-bounded response and high-dimensional correlated covariates of mixed types. The proposed method makes full use of all available data through one or several linear combinations of the covariates without losing information from the data. The only assumption made is that the response variable follows a beta distribution, and no additional assumption is required. The resulting estimators are consistent and efficient. The proposed method is illustrated in simulation studies and is demonstrated in a real-life data application. The semiparametric approach contributes to the sufficient dimension reduction literature for its novelty in investigating double-bounded response, which is absent in the current literature. A new tool is also provided for data practitioners to analyze the association between a popular unit interval response and mixed types of high-dimensional correlated covariates.

C1294: **Microbial co-abundance network with community detection**
Presenter: **Yan Li**, Auburn University, United States

Microbial network analysis holds the potential to decipher complicated ecological interactions between microbes and, therefore, enhance the understanding of microbial functionality. However, existing methods are inadequate to accommodate the compositionality of microbiome data or extract concise and insightful dependence structures from high-dimensional data. A new network framework with structural constraints to the microbiome data analysis is proposed based on the notion of Laplacian-constrained Gaussian graphical models. The proposed Laplacian constraint model provides a perfect fit for the unique features of centered log-ratio transformed compositional data with a linear constraint. Under the proposed framework, zero-sparsity and nuclear-norm regularization methods and an efficient computation algorithm are developed to achieve a highly interpretable co-abundance network with biologically meaningful community separation. Theoretical properties are explored for the proposed model. The efficacy of the proposed methods is demonstrated in extensive simulation studies and real gut microbiome studies.

C1266: **Debiased high-dimensional regression calibration for errors-in-variables log-contrast models**
Presenter: **Tianying Wang**, Colorado State University, United States

Motivated by the challenges in analyzing gut microbiome and metagenomic data, the aim is to tackle the issue of measurement errors in high-dimensional regression models that involve compositional covariates. A pioneering effort is made to conduct statistical inference on high-dimensional compositional data affected by mismeasured or contaminated data. A calibration approach tailored to the linear log-contrast model is introduced. Under relatively lenient conditions regarding the sparsity level of the parameter, the asymptotic normality of the estimator for inference is established. Numerical experiments and an application in microbiome study have demonstrated the efficacy of the high-dimensional calibration strategy in minimizing bias and achieving the expected coverage rates for confidence intervals. Moreover, the potential application of the proposed methodology extends well beyond compositional data, suggesting its adaptability for a wide range of research contexts.

CO317 Room S-1.27 SELECTED TOPICS IN STATISTICAL MACHINE LEARNING Chair: Katherine Thompson

C0624: **Dependence structure estimation from incomplete data**
Presenter: **Giuseppe Vinci**, University of Notre Dame, United States

Modern scientific multivariate data sets are often incomplete and corrupted by noise. Implementing statistical methods is very challenging in these situations, which are common in numerous science fields, including neuroscience, genomics, astronomy, and forensic science. In particular, the estimation of the covariance matrix and related objects, such as conditional dependence graphs, is nearly impossible when the data are structurally incomplete. Novel methods for the high-dimensional estimation of covariance matrices and graphical models from incomplete and corrupted data are presented. Two families of approaches are discussed: factor analysis methods and sparse precision matrix estimation methods. Theory and methods are presented for traditional random vectors and also in the framework of multivariate extreme values. Simulation studies and applications to real data are also presented.

C0843: **Exploring Bayesian learning with heterogeneous image sources: From non-parametric models to deep learning architectures**
Presenter: **Rajarshi Guhaniyogi**, Texas A & M university, United States
Co-authors: Aaron Scheffler

Various disciplines have recently encountered objects (such as tensors and networks) from multiple sources exhibiting diverse scales. These structured datasets offer a wealth of shared information among objects, which could potentially unveil crucial scientific insights through sophisticated analysis. Hierarchical Bayesian modeling approaches have the potential to effectively amalgamate information from heterogeneous objects using joint prior structures, thereby enabling comprehensive uncertainty estimation in inference. However, the adoption of this approach has been limited by the absence of robust Bayesian methods that yield precise inference, along with challenges in computation and theory. Innovative regression frameworks are introduced with an object outcome and heterogeneous object predictors encompassing both Bayesian non-parametric models and interpretable deep learning architecture to address these inferential challenges simultaneously. To validate the approach, empirical evidence is provided, particularly by analyzing high-impact multi-modal brain imaging datasets in collaboration with neuroscientists.

C1139: **Machine learning methods: Probability of correct model selection using R^2 or AIC**
Presenter: **Katherine Thompson**, University of Kentucky, United States

Although recent attention has focused largely on improving predictive models, less consideration has been given to the prevalence of incorrect models selected by traditional statistical methods. The difficulty in choosing a scientifically correct model is quantified through theoretical and simulation work. Furthermore, the performance of traditional model selection techniques is compared with that of the feasible solutions algorithm, a recent machine learning method. Specifically, when data sets contain large numbers of explanatory variables, the model with the highest R^2 (or adjusted R^2) or lowest AIC is often not the scientifically correct model that produced the data, suggesting that traditional model selection techniques that rely on these criteria may be inappropriate. It starts with the derivation of the probability of choosing the scientifically correct model in data sets as a function of regression model parameters when using R^2 or AIC. Next, simulation results show that these traditional model selection criteria are outperformed by the feasible solutions algorithm, a machine learning method that produces multiple candidate models for researchers' consideration. Lastly, these results are demonstrated through the analysis of a National Health and Nutrition Examination Survey data set.

C1605: **A progressive latent space learning framework for improving the robustness of information bottleneck**
Presenter: **Anirban Samaddar**, Argonne National Laboratory, United States

The information bottleneck framework provides a systematic approach to learning representations that compress nuisance information in the input and extract semantically meaningful information about predictions. However, the choice of a prior distribution that fixes the dimensionality across all the data can restrict the flexibility of this approach for learning robust representations. A novel sparsity-inducing spike-slab categorical prior is presented that uses sparsity as a mechanism to provide the flexibility that allows each data point to learn its own dimension distribution. In addition, it provides a mechanism for learning a joint distribution of the latent variable and the sparsity, and hence, it can account for the complete uncertainty in the latent space. Through a series of experiments using in-distribution and out-of-distribution learning scenarios on the MNIST, CIFAR-10, and ImageNet data, it is shown that the proposed approach improves accuracy and robustness compared to traditional fixed-dimensional priors, as well as other sparsity induction mechanisms for latent variable models proposed in the literature.

CO104 Room Auditorium NON-GAUSSIAN AND NONCAUSAL TIME SERIES Chair: Joann Jasiak

C0177: **Mixed causal-noncausal count process**
Presenter: **Yang Lu**, Concordia University, Canada
Co-authors: Jian Pei

Recently, a study introduced a class of (Markov) noncausal count processes. These processes are obtained by time-reverting a standard count process (such as INAR(1)) but have quite different dynamic properties. In particular, they can feature bubble-type phenomena, which are epochs of steady increase followed by sharp decreases. This is in contrast to usual INAR and INGARCH type models, which only feature "reverse bubbles", which are epochs of sharp increase followed by steady decreases. In practice, however, in many datasets, sudden jumps or crashes are rare, and instead, it is more frequent to observe epochs of steady increase or decrease. In practice, count time series data can feature both bubbles and reverse bubbles. This raises the question of what model to use if the modeler cannot discriminate a priori between the causal and noncausal models. The mixed causal-noncausal integer-valued autoregressive (MINAR(1,1)) process is introduced by superposing a causal and a noncausal INAR(1) process, sharing the same sequence of error terms. This process inherits some key properties from the noncausal INAR(1), such as the bi-modality of the predictive distribution and the irreversibility of the dynamics, while at the same time allowing different accumulation and burst speeds for the bubble. A GMM estimation method is proposed, its finite sample performance is investigated, testing procedures are developed, and the methodology is applied to stock transaction data.

C0379: **Extracting efficient prices using intrinsic time information**
Presenter: **Roxana Halbleib**, University of Freiburg, Germany
Co-authors: Lukas Schmidt-Engelbertz

A comprehensive framework is presented designed to extract efficient price processes on financial markets, acknowledging and mitigating the impact of market microstructure noise inherent in high-frequency financial data. The framework comprises three pivotal stages: sampling, filtering, and optimizing. In the sampling stage, various frequencies and schemes are explored, both calendar and intrinsic time-based, to capture noisy ultra-high frequency prices effectively. Following the sampling stage, the methodology employs a Kalman filter in the filtering stage. This filter is utilized across multiple window sizes, aiming to extract the fundamental price process while minimizing the impact of market microstructure noise. The optimization stage uses a non-linear optimization approach to minimize the autocorrelation left, thus aiming to obtain a price process that exhibits martingale properties. The results demonstrate the success of this methodology in removing significant autocorrelation present in observed high-frequency financial prices. Furthermore, the approach offers adaptability and versatility, permitting methodological substitutions at each stage. This flexibility empowers researchers to tailor the framework to better suit the specific nature of their data or underlying models.

C0429: **GCov-based Portmanteau test**
Presenter: **Aryan Manafi Neyazi**, York University, Canada
Co-authors: Joann Jasiak

The purpose is to study nonlinear serial dependence tests for non-Gaussian time series and residuals of dynamic models based on Portmanteau statistics involving nonlinear autocovariances. A new NLSD test with an asymptotic chi-square distribution is introduced to test nonlinear serial dependence in time series. This test is inspired by the generalized covariance (GCov) residual-based specification test, recently proposed as a diagnostic tool for semi-parametric dynamic models with i.i.d. non-Gaussian errors. It has a chi-square distribution when the model is correctly specified and estimated by the GCov estimator. It is extended by introducing a GCov bootstrap test for residual diagnostics when the model is estimated by a different method, such as the maximum likelihood estimator under a parametric assumption on the error distribution. The GCov specification test is reviewed, and new asymptotic results are derived under local alternatives for testing hypotheses on the parameters of a semi-parametric model. A simulation study shows that the tests perform well in applications to mixed causal-noncausal univariate and multivariate autoregressive models. The GCov specification test is used to assess the fit of a mixed causal-noncausal model of aluminum prices with locally explosive patterns, i.e. bubbles and spikes between 1990 and 2023.

C0451: **Forecasting extreme trajectories using semi-norm representations**
Presenter: **Arthur Thomas**, Paris Dauphine University - PSL, France
Co-authors: Gilles De Truchis, Sebastien Fries

For a two-sided stable moving average, the conditional distribution of future paths is studied, given a piece of observed trajectory when the process is far from its central values. Under this framework, vectors of the form $Xt = (Xt - m, ..., Xt, Xt + 1, ..., Xt + h)$, $m > 0$, $h > 1$, are multivariate stable, and the dependence between the past and future components is encoded in their spectral measures. A new representation of stable random vectors on unit cylinder sets for an adequate semi-norm is proposed to describe the tail behaviour of vectors Xt when only the first m+1 components are assumed to be observed and large in norm. Not all stable vectors admit such a representation, and (Xt) must be anticipative enough for Xt to admit one. The conditional distribution of future paths can then be explicitly derived using the regularly varying tails property of stable vectors, which has a natural interpretation in terms of pattern identification. Through Monte Carlo simulations, procedures are developed to forecast crash probabilities and crash dates and demonstrate their finite sample performances. Probabilities and reversal dates of El Nio and La Nia occurrences are estimated as an empirical illustration.

CO131 Room K0.16 NOVEL INFERENCE AND MODELING ON NETWORK DATA AND APPLICATIONS Chair: Wen Zhou

C0166: **Asymptotic failure of peer effects in network regression models**
Presenter: **Keith Levin**, University of Wisconsin, United States

Network autoregressive models seek to model peer effects, such as contagion and interference, in which node-level responses or behaviors may influence one another. These models are frequently deployed by practitioners in sociology and econometrics, typically in the form of linear-in-means models, in which node-level covariates and local averages of responses are used as predictors. In highly structured networks, previous work has shown that peer effects in linear-in-means models are collinear with other regression terms and thus cannot be estimated, but this collinearity is widely believed to be ignorable, as peer effects are typically identified in empirical networks. A concerning negative result is shown: Under linear-in-means models when node-level covariates are independent of network structure, peer effects become increasingly collinear with other regression terms as the network size (i.e., number of nodes) grows and are inestimable asymptotically. A narrow positive result is also shown: Under certain latent space network models, some peer effects remain identified as the network size grows, albeit under rather stringent conditions. The results suggest that linear models for peer effects are appropriate in far fewer settings than was previously believed.

C0314: **Preferential latent space models for networks with textual edges**
Presenter: **Emma Jingfei Zhang**, Emory University, United States

Many real-world networks contain rich textual information at the edges, such as email networks where an edge between two nodes is an email exchange. Other examples include co-author networks and social media networks. The useful textual information carried in the edges is often discarded in most network analyses, resulting in an incomplete view of the relationships between nodes. The aim is to propose representing the text document between each pair of nodes as a vector counting the appearances of keywords extracted from the corpus and introduce a new and flexible preferential latent space network model that can offer direct insights into how contents of the textual exchanges modulate the relationships between nodes. Identifiability conditions are established for the proposed model, and model estimation is tackled with a computationally efficient projected gradient descent algorithm. The non-asymptotic error bound is further derived from the estimator from each step of the algorithm. The efficacy of the proposed method is demonstrated through simulations and an analysis of the Enron email network.

C0433: **Simultaneous estimation of connectivity and dimensionality in samples of networks**
Presenter: **Joshua Cape**, University of Wisconsin, Madison, United States
Co-authors: Wenlong Jiang, Jesus Arroyo, Christopher McKennan

A method is proposed to simultaneously estimate a latent connectivity matrix and its embedding dimensionality (rank) after first pre-estimating the number of communities and node cluster memberships. The method is formulated as a convex optimization problem and solved using an alternating direction method of multipliers algorithm. Estimation error bounds are established under the Frobenius norm and nuclear norm for settings in which observable networks have blockmodel structure. Numerical studies empirically demonstrate the accuracy of our method even when node block memberships are imperfectly recovered. When exact membership recovery is possible and dimensionality is much smaller than the number of communities, the method outperforms averaging-based methods for estimating connectivity and dimensionality. Analysis of a primate brain dataset demonstrates that posited connectivity is not necessarily full rank in practice, illustrating the need for flexible methodology.

C1162: **Regression discontinuity designs under interference**
Presenter: **Laura Forastiere**, Yale University, United States

Interference takes place whenever a treatment on one unit affects the outcome of another unit, and such a phenomenon can also occur in regression discontinuity designs (RDDs). For instance, in conditional cash transfer programs for education, the eligibility to the program and the potential receipt of cash transfers may affect eligible children's schooling choices, which in turn may influence schooling choices of their peers. In this setting, assignment to the individual treatment and to the spillover exposure, which incorporates through a mapping function the exposure to the treatment of interfering units (e.g., friends, classmates), is determined by a unit's score and the scores of other interfering units. Unlike the standard RDD, the presence of spillover exposure to other units may give rise to complex, multidimensional boundaries on a multidimensional score space. A method is provided to characterize such boundaries for a broad class of exposure mapping functions and derive generalized continuity assumptions to identify the proposed causal estimands. Next, an estimation method that can handle multidimensional and potentially heterogeneous multi-scores, including complex dependencies, is developed. Finally, the proposed methodology is applied to the PROGRESA/Oportunidades data to estimate the direct and indirect effects of receiving cash transfers on children's schooling attendance.

CO083 Room K0.18 STATISTICAL METHODS IN WEATHER FORECASTING I Chair: Bastien Francois

C1265: **Improving probabilistic forecasts of extreme winds by training post-processing models with weighted scoring rules**
Presenter: **Jakob Wessel**, University of Exeter, United Kingdom
Co-authors: Christopher Ferro, Gavin Evans, Frank Kwasniok

Accurate forecasts of extreme wind speeds are of high importance for many applications. Such forecasts are usually generated by ensembles of numerical weather prediction (NWP) models, which, however, can be biased and have errors in dispersion, thus necessitating the application of statistical post-processing techniques. The aim is to improve statistical post-processing models for probabilistic predictions of extreme wind speeds. It is done by adjusting the training procedure used to fit ensemble model output statistics (EMOS) models - a commonly applied post-processing technique - and propose estimating parameters using the so-called threshold-weighted continuous ranked probability score (twCRPS), a proper scoring rule that places special emphasis on predictions over a threshold. It is shown that training using the twCRPS leads to improved extreme event performance of post-processing models for a variety of thresholds. A distribution body-tail trade-off is found where improved performance for probabilistic predictions of extreme events comes with worse performance for predictions of the distribution body. However, strategies to mitigate this trade-off are introduced based on weighted training and linear pooling. Finally, some synthetic experiments are considered to explain the impact of training on the twCRPS. The results enables researchers and practitioners alike to improve the performance of probabilistic forecasting models for extremes and other events of interest.

C1289: **Fair logarithmic score for multivariate Gaussian forecasts**
Presenter: **Sandor Baran**, University of Debrecen, Hungary
Co-authors: Martin Leutbecher

The verification of multivariate probabilistic forecasts is concerned with evaluating joint probability distributions of vector quantities, such as a weather variable at multiple locations or a wind vector, using scoring rules. The focus is on the strictly proper logarithmic score, which can be applied to ensemble weather forecasts sampled from a specific distribution. Under the assumptions of multivariate normality with mean and covariance matrix given by the ensemble mean and ensemble covariance, respectively, the dependence of the logarithmic score is derived from the ensemble size. It permits estimating the score in the limit of infinite ensemble size from a small ensemble and thus produces a fair logarithmic score. Using ensemble forecasts of vectors consisting of various combinations of upper-air weather variables, the usefulness of the ensemble size adjustments is demonstrated when multivariate normality is only an approximation. It is shown that fair logarithmic scores of ensembles with different cardinalities are very close, in contrast to their unadjusted counterparts, which decrease considerably with ensemble size.

C1337: **Probabilistic post-processing of wind speed forecasts with explicit modeling of time dependencies**
Presenter: **Katharina Klein**, Utrecht University, Netherlands
Co-authors: Sjoerd Dirksen, Maurice Schmeits, Kirien Whan

Post-processing NWP forecast data is essential for removing biases and obtaining better-calibrated probabilistic forecasts. Post-processing multiple lead times simultaneously is particularly challenging due to the inherent temporal dependencies, which classical approaches deal with by using a two-step procedure: lead times are first processed independently, and (empirical) copula methods are subsequently applied to restore dependencies. The proposed ARMOS model, a generalization of the widely used ensemble model output statistics (EMOS) for estimating marginal distributions, can circumvent the second step by incorporating temporal dependencies explicitly. It exploits the autoregressive property of forecast errors and yields a multivariate probability distribution for the given weather variable. The ARMOS model is applied to deterministic forecasts from KNMI's Harmonie-Arome model in order to obtain a multivariate parametric distribution for wind speed forecasts from initialization time to 48 hours ahead. Compared to a state-of-the-art two-step approach (EMOS adapted to deterministic forecasts and paired with Schaake-Shuffle), ARMOS shows similar or better performance on different multivariate and univariate evaluation methods. The model is thus effective in post-processing NWP forecasts for multiple lead times without the need to use empirical copula methods.

C1373: **Incorporating climatological constraints into statistical models to improve the post-processing of extremes**
Presenter: **Bastien Francois**, KNMI, Netherlands
Co-authors: Harun Kivril, Maurice Schmeits, Kirien Whan, Philippe Naveau

Extreme events, such as extreme wind gusts, can generate huge impacts on society, and anticipating them is essential for taking preventive measures. Ensemble forecasts exhibit biases and under-dispersion and have to be calibrated using observations before being used. Several statistical post-processing methods have, therefore, been developed and applied. Among them, some are non-parametric and are thus not able to predict beyond the largest value observed during training. To overcome this problem, parametric distributions such as those from extreme value theory (EVT) can be fitted on post-processed outputs to allow extrapolation. However, depending on the outputs, the extrapolation may sometimes be insufficient. A new method is proposed to enable the extrapolation of post-processed outputs by forcing the fitted EVT distributions to have an upper bound aligned with the maximum of the climatological records at the stations. The proposed method is applied to forecasts of 6-hourly maximum wind gusts from 2021 to 2024 over the Netherlands using the ECMWF-IFS ensemble data. Results are compared against several state-of-the-art post-processing methods. The proposed algorithm shows skill improvements for wind gust extremes depending on lead times, stations and thresholds while maintaining performances for intermediate values. It encourages further research on adding better constraints to methods for the post-processing of extremes.

CO410 Room K0.19 STATISTICAL AND COMPUTATIONAL METHODS FOR LONGITUDINAL AND SURVIVAL DATA Chair: Panpan Zhang

C0714: **Bayesian modelling and inference for finite populations from process-based superpopulations**
Presenter: **Sudipto Banerjee**, UCLA, United States

The purpose is to offer some perspectives on Bayesian inference for finite population quantities when the units in the population are assumed to exhibit complex dependencies. More specifically, Bayesian models are developed when the finite population units are assumed to be realizations of a spatial process. With an overview of Bayesian hierarchical models, including some yielding design-based Horvitz-Thompson estimators, dependence is introduced in finite populations, and inferential frameworks are set out for ignorable and nonignorable responses. Multivariate dependencies using graphical models and spatial processes are discussed, and some salient features of two recent analyses of spatially oriented finite populations are presented.

C0731: **Data harmonization via regularized nonparametric mixing distribution estimation**
Presenter: **Kwun Chuen Gary Chan**, University of Washington, United States
Co-authors: Steven Wilkins-Reeves, Yen-Chi Chen

Data harmonization is the process of developing an equivalence between two measurements of a common domain. The problem is motivated by dementia research in which multiple neuropsychological tests have been used in practice to measure the same underlying cognitive ability, such as memory or attention. This statistical problem is connected to mixing distribution estimation, which is common in empirical Bayes approaches. A nonparametric latent trait model is introduced and studied; a method is developed that enforces the uniqueness of the regularized maximum likelihood estimator, showing how a nonparametric EM algorithm will converge weakly to its maximizer and its superior computational efficiency to off-the-shelf solvers is illustrated. The method is applied to the National Alzheimer's Coordination Center uniform data set, and it is shown that the method can be used to convert between score measurements and account for the measurement error. It is shown that this method outperforms standard techniques commonly used in dementia research.

C0957: **Cox regression model with auxiliary endpoints accounting for left truncation, complex censoring, and missing data**
Presenter: **Sharon Xie**, University of Pennsylvania, United States
Co-authors: Yidan Shi

The time-to-event analysis is a widely employed approach for modelling disease progression data. However, obtaining the true survival endpoint, such as the age at which cerebrospinal fluid biomarkers for Alzheimer's disease become abnormal, can be costly, invasive, and contingent on participants' cognitive status. As a result, it is often accessible only to a small subset of participants, which can affect estimation accuracy and efficiency. An auxiliary event, which is less costly or invasive but may measure the true event with error, is often available. Additionally, delayed entry in observational studies can result in left truncation, which further complicates analyses. Since the auxiliary event is not the primary focus, it often suffers from left censoring due to study design. A likelihood-based method and an E-M algorithm for Cox regression models that incorporate the auxiliary endpoint while accounting for left truncation, complex censoring scenarios, and auxiliary-event-dependent missingness are proposed. The method improves efficiency and corrects for bias compared to complete case analysis. It is demonstrated that the proposed regression coefficient estimator is consistent and asymptotically normally distributed. The performance of the method is assessed in finite sample scenarios through simulation studies. Finally, the proposed method is illustrated using the Alzheimer's disease neuroimaging initiative (ADNI) study.

C1062: **Predictive partly conditional model for longitudinal outcomes in the presence of informative dropout and death**
Presenter: **Dandan Liu**, Vanderbilt University Medical Center, United States

Assessing time-dependent risk factors in relation to the risk of disease progression is challenging yet important, especially for chronic diseases with slow progression. An important consideration when modelling outcomes related to ageing is the potential for dropout in the study or death prior to the disease occurring. The predictive partly conditional model (PPCM) is extended to characterize disease progression at time t with longitudinal outcomes in the presence of time-dependent covariates at time $s(s < t)$ when informative death and dropout are present. Inverse probability weighting is adopted to separately account for the probability of death and dropout with the generalized estimating equation approach and establish the conditions for consistency. Extensive simulation studies are conducted to assess the properties of the proposed model and demonstrate the implementation of the proposed model using existing statistical software. Application to data from the National Alzheimer's Coordinating Center will be conducted to illustrate this method.

CO117 Room K0.20 DISTANCE BASED METHODS IN MODEL SPECIFICATION TESTING AND SELECTION Chair: Bojana Milosevic

C0388: **Comparing mixing data**
Presenter: **Maria Dolores Jimenez-Gamero**, Universidad de Sevilla, Spain
Co-authors: Virtudes Alba-Fernandez, Francisca Jimenez-Jimenez

The problem of testing the equality of two populations based on independent samples from each of them (the two-sample problem) is a classic one in statistics. Most existing tests assume that all data are categorical or numerical. The mixed data case (some components of the vector to be compared are categorical, and some are numerical) has received less attention in the statistical literature. A new test for comparing mixed data is proposed and studied. The null distribution of the test statistic can be approximated by permutation or bootstrap. The new test is consistent against fixed alternatives. The performance of the test in finite samples is studied by simulation. The proposal is applied to a real data set.

C1420: **Distance covariance for random fields**
Presenter: **Muneya Matsui**, Nanzan University, Japan

An independence test is studied based on distance correlation for random fields (X,Y). The situations are considered when (X,Y) is observed on a lattice with equidistant grid sizes and when (X,Y) is observed at random locations. The asymptotic theory is provided for the sample distance correlation in both situations, and bootstrap consistency is shown. The latter fact allows one to build a test for independence of X and Y based on the considered discretization of these fields. The performance of the bootstrap test is illustrated in a simulation study involving fractional Brownian and infinite variance stable fields. The independence test is applied to Japanese meteorological data, which are observed over the entire area of Japan.

C1272: **Sequential change-point detection for skew normal distribution**
Presenter: **Wei Ning**, Bowling Green State University, United States

A modified max cumulative sum (CUSUM) procedure is proposed to detect changes in the parameters of the skew's normal distribution. The corresponding false alarm frequency and the post-change detection delay are investigated. Asymptotic behaviors of detection delay and theoretical optimality of the detection procedure have been established. Simulations have been conducted to show the performance of the proposed method and compare it to the other existing methods, including CUSUM. Real data are given to illustrate the detection procedure.

C1091: **A Wasserstein perspective of Vanilla GANs**
Presenter: **Lea Kunkel**, Karlsruhe Institute of Technology, Germany
Co-authors: Mathias Trabs

The empirical success of generative adversarial networks (GANs) caused an increasing interest in theoretical research. The statistical literature is mainly focused on Wasserstein GANs and generalizations thereof, which especially allow for good dimension reduction properties. Statistical results for Vanilla GANs, the original optimization problem, are still rather limited and require assumptions such as smooth activation functions and equal dimensions of the latent space and the ambient space. To bridge this gap, a connection is drawn from the distance approximated by Vanilla GANs to the Wasserstein distance. By doing so, existing results for Wasserstein GANs can be extended to Vanilla GANs. In particular, an oracle inequality for Vanilla GANs in Wasserstein distance is obtained. The assumptions of this oracle inequality are designed to be satisfied by network architectures commonly used in practice, such as feedforward ReLU networks. By providing a quantitative result for the approximation of a Lipschitz function by a feedforward ReLU network with bounded Hoelder norm, a rate of convergence Vanilla GANs is concluded as estimators of the unknown probability distribution.

CO270 Room K0.50 DESIGN AND ANALYSIS OF COMPLEX EXPERIMENTS (VIRTUAL) Chair: MingHung Kao

C1192: **Fast approximation of Shapley values through fractional factorial designs**
Presenter: **Wei Zheng**, University of Tennessee, United States
Co-authors: Zheng Zhou, Robert Mee, Herbert Hamers

Shapley value is a well-known concept in cooperative game theory that provides a fair way to distribute revenues or costs to each player. Recently, it has been widely applied in various fields, such as data science, marketing, and genetics. However, the computation of the Shapley value is an NP-hard problem. For a cooperative game with n players, calculating Shapley values for all players requires calculating the value function for 2^n different coalitions, which makes it infeasible for a large n. A fast approximation approach is proposed for Shapley values based on fractional factorial designs. Under certain conditions, the approach can obtain true Shapley values by calculating values of fewer than $4n^2-4$ different coalitions. In general, highly accurate approximations of Shapley values can also be obtained by calculating values of additional $O(n^2)$ different coalitions. Multiple simulations and real case examples demonstrate that, with equivalent computational cost, the method provides significantly more accurate approximations compared with several popular methods.

C0461: **Optimal designs with multiple correlated responses for experiments with mixtures**
Presenter: **Hsiang-Ling Hsu**, National University of Kaohsiung, Taiwan

A mixture experiment within the (q-1)-dimensional probability simplex is a specific experimental setup in which the q factors are non-negative and adhere to the sum of all factors equals one. The issue of the optimal approximate designs is investigated with the k-correlated response mixture experimental models. In the multiple correlated response mixture models, the improvement design class is explored, known as the complete class, in relation to the Kiefer ordering for a given design. Based on the complete class results, the properties of optimal designs are delved into for multiresponse models using the well-established equivalence theorem. For specific multiresponse model settings under the D- or A-optimal design criteria, the optimal results can reduce multiresponse experimental design problems to single-response experimental design problems. An illustrative example showcasing optimal designs for correlated response mixture experimental models is presented.

C0462: **An efficient approach for identifying important biomarkers for biomedical diagnosis**
Presenter: **Jing-Wen Huang**, Academia Sinica, Taiwan
Co-authors: Yan-Hong Chen, Frederick Kin Hing Phoa, Yan-Han Lin, Shau Ping Lin

The challenges associated with biomarker identification are explored for diagnosis purposes in biomedical experiments, and a novel approach is proposed, inspired by the analysis of supersaturated designs, to handle the above challenging scenario via the generalization of the Dantzig selector. To improve the efficiency of the regularization method, a transformation is introduced from an inherent nonlinear programming due to its nonlinear link function into a linear programming framework under a reasonable assumption on the logistic probability range. The use of the method is illustrated in an experiment with binary response, showing superior performance on biomarker identification studies when compared to their conventional analysis. The proposed method does not merely serve as a variable/biomarker selection tool; its ranking of variable importance provides valuable reference information for practitioners to reach informed decisions regarding the prioritization of factors for further investigations.

C0729: **Optimal next stage designs for sparse longitudinal data**
Presenter: **MingHung Kao**, Arizona State University, United States

The focus is on optimal designs for collecting informative sparse functional/longitudinal data to precisely predict the trajectories of individual random curves. Constructing such designs typically requires knowledge of unknown quantities of the model. For this issue, previous studies considered locally optimal designs by replacing these quantities with their estimates from the pilot study. However, the information on the design used for collecting the pilot data is completely discarded when developing the design for a future study. To address this drawback, a multi-stage design approach is proposed to adapt the optimal design for the next study based on the design for the pilot study. A new optimality criterion is developed to select a design that gives a precise curve prediction for the next study and improves the curve prediction in the pilot study. Useful theories and computational approaches to facilitate the identification of optimal designs are also provided.

CO080 Room K2.31 (Nash Lec. Theatre) ADVANCES IN CAUSAL INFERENCE Chair: Luke Keele

C0423: **Local instrumental variable curves without positivity assumptions**
Presenter: **Luke Keele**, University of Pennsylvania, United States

Instrumental variables have become a popular study design for the estimation of treatment effects in the presence of unobserved confounders. In the canonical instrumental variables design, the instrument is a binary variable, and most extant methods are tailored to this context. In many settings, however, the instrument is a continuous measure. While standard estimation methods can be applied to continuous instruments, these methods require strong functional form assumptions. Recent work has developed more flexible methods for the estimation when the instrument is continuous. However, these methods require an assumption known as positivity that is unlikely to hold for many applications. Doubly robust estimators are developed for continuous instruments that do not require a positivity assumption. The methods use a stochastic dynamic intervention framework that considers a range of intervention distributions absolutely continuous with respect to the observed distribution of the instrument. Empirical process theory and sample splitting are used to derive asymptotic properties of an estimator under weak conditions. Estimation methods also allow for the use of nonparametric estimators for the nuisance functions. The methods are evaluated via simulation and demonstrate their feasibility using an application on the effectiveness of surgery for specific emergency conditions.

C0498: **Evaluating the validity and robustness of instrumental-variable analyses**
Presenter: **Dean Knox**, UPenn Wharton, United States
Co-authors: Luke Keele, Guilherme Duarte, Kai Cooper, Jonathan Mummolo, Kennedy Mattes

Instrumental-variable (IV) designs are widely used across numerous fields to estimate causal effects when the relationship between treatment and outcome is confounded, exploiting as-if randomized encouragements that nudge units into treatment. The validity of these designs rests on several assumptions that are often regarded as difficult to test, including monotonicity, the assumption that no units defy encouragement and exclusion, and the assumption that the instrument does not directly affect the outcome. Using newly derived analytic results and recent advances in automated partial identification, a range of falsification tests and sensitivity analyses are presented that empirically evaluate the validity of these assumptions and the robustness of inferences to their violations. Replicating and extending published examples, it is shown how the techniques are flexible to the idiosyncrasies of applied settings by sharply bounding causal estimands in situations where key IV assumptions are believed to fail.

C0540: **Effect aliasing in observational studies**
Presenter: **Jose Zubizarreta**, Harvard University, United States
Co-authors: Paul Rosenbaum

In experimental design, aliasing of effects occurs in fractional factorial experiments, where certain low-order factorial effects are indistinguishable from certain high-order interactions: low-order contrasts may be orthogonal to one another, while their higher-order interactions are aliased and not identified. In observational studies, aliasing occurs when certain combinations of covariates, e.g., time period and various eligibility criteria for treatment, perfectly predict the treatment that an individual will receive, so a covariate combination is aliased with a particular treatment. In this situation, when a contrast among several groups is used to estimate a treatment effect, collections of individuals defined by contrast weights may be balanced with respect to summaries of low-order interactions between covariates and treatments but necessarily not balanced with respect to summaries of high-order interactions between covariates and treatments. A theory of aliasing is developed in observational studies, illustrating that theory in an observational study whose aliasing is more robust than conventional difference-in-differences, and a new form of matching is developed to construct balanced confounded factorial designs from observational data.

C1061: **Simulating data from marginal structural models for a survival time outcome**
Presenter: **Shaun Seaman**, University of Cambridge, United Kingdom
Co-authors: Ruth Keogh

Marginal structural models (MSMs) are often used to estimate the causal effects of treatments on survival time outcomes from observational data when time-dependent confounding may be present. They can be fitted using, e.g., inverse probability of treatment weighting (IPTW). It is important to evaluate the performance of statistical methods in different scenarios, and simulation studies are a key tool for such evaluations. In such simulation studies, it is common to generate data in such a way that the model of interest is correctly specified, but this is not always straightforward when the model of interest is for potential outcomes, as is an MSM. Methods have been proposed for simulating MSMs for a survival outcome, but these methods impose restrictions on the data-generating mechanism. A method is proposed to overcome these restrictions. The MSM can be, for example, a marginal structural logistic model for a discrete survival time or a Cox or additive hazards MSM for a continuous survival time. The hazard of the potential survival time can be conditional on baseline covariates, and the treatment variable can be discrete or continuous. The use of the proposed simulation algorithm is illustrated by carrying out a brief simulation study. The coverage of confidence intervals calculated in two different ways is compared for causal effect estimates obtained by fitting an MSM via IPTW.

CO228 Room K2.40 HIGH-DIMENSIONAL TIME SERIES AND DATA INTEGRATION Chair: Vladas Pipiras

C0649: **Discovering common structures across high-dimensional factor models**
Presenter: **Marie Duker**, FAU Erlangen, Germany

A sequential testing procedure is proposed to uncover common structures across multiple high-dimensional factor models. The test is motivated by observing data from multiple individuals, which can be modeled through factor models that potentially share information encoded in their respective loading matrices. The introduced sequential procedure allows testing whether these loading matrices are identical up to a rotational change or if only a partial set of column vectors is shared across individuals. The theoretical results cover the asymptotic behavior of the test statistic, supported by a simulation study demonstrating promising empirical test size and power. Finally, the method is applied to investigate the relationship between multiple individuals with anxiety disorder.

C0723: **Data integration via analysis of subspaces (DIVAS)**
Presenter: **Jan Hannig**, University of North Carolina at Chapel Hill, United States
Co-authors: Quoc Tran-Dinh, Jack Prothero, Andrew Ackerman, Meilei Jiang, Steve Marron

Modern data collection in many data paradigms, including bioinformatics, often incorporates multiple traits derived from different data types (i.e. platforms). This data is called multi-block, multi-view, or multi-omics data. The emergent field of data integration develops and applies new methods for studying multi-block data and identifying how different data types relate and differ. One major frontier in contemporary data integration research is a methodology that can identify partially-shared structure between sub-collections of data types. A new approach is presented: Data integration via analysis of subspaces (DIVAS). DIVAS combines new insights in angular subspace perturbation theory with recent developments in matrix signal processing and convex-concave optimization into one algorithm for exploring partially-shared structure. Based on principal angles between subspaces, DIVAS provides built-in inference on the results of the analysis and is effective even in high-dimension-low-sample-size (HDLSS) situations.

C0848: **Characterizing heterogeneous dynamics in multiple-subject multivariate time series**
Presenter: **Zachary Fisher**, Penn State University, United States

Heterogeneity is a ubiquitous and defining feature of human behavior. At the same time, how best to characterize and intervene on heterogeneous processes remains a critical open question. Part of the difficulty in addressing heterogeneity lies in the fact that individuals differ from one another in complex and meaningful ways, and at the same time, dynamics within individuals themselves often evolve and adapt across time and context. Recently, a study introduced the multi-VAR framework for simultaneously modeling multiple-subject multivariate time series characterized by common and individualizing features using penalized estimation. This approach differs from many popular modeling approaches for multiple-subject time series in that both qualitative and quantitative differences in a large number of individual-level dynamics are well-accommodated. Extensions of the original multi-VAR approach are presented to accommodate nonstationary series by allowing individual-level dynamics to vary over time.

C0968: **A regularized low tubal-rank model for high-dimensional time series data**
Presenter: **Samrat Roy**, Indian Institute of Management Ahmedabad, India

High-dimensional time series analysis has diverse applications in macroeconomics and finance. Recent factor-type models employing tensor-based decompositions prove to be computationally involved due to the non-convex nature of the underlying optimization problem, and they do not capture the underlying temporal dependence of the latent factor structure. The concept of tubal rank is leveraged, and a matrix-valued time series model is developed, which first captures the temporal dependence in the data, and then the remainder signals across the time points are decomposed into two components: a low tubal rank tensor representing the baseline signals, and a sparse tensor capturing the additional idiosyncrasies in the signal. The issue of identifiability of various components is addressed in the model, and a scalable alternating block minimization algorithm is subsequently developed to solve the convex regularized optimization problem for estimating the parameters. Finite sample error bounds are provided under high dimensional scaling for the model parameters.

CO396 Room K2.41 SURVIVAL ANALYSIS: TRUNCATED DATA Chair: Takeshi Emura

C0208: **Left-truncated durations: Theoretical review and empirical applications**
Presenter: **Rafael Weissbach**, Univerisity of Rostock, Germany
Co-authors: Eric Scholz

Left-truncated are statistical units in a panel data set when (i) Units born before the first wave belong to the population, (ii) Death of a population unit before the first wave is possible, and (iii) The birthdate of every observed unit is known if the model is not age-homogeneous. Filtrations model the missing data and enable circumventing the unknown number of truncated units by using conditional likelihood. Proving the martingale property of the process of counting events of interest after compensation enables the use of martingale limit theorems to derive standard errors and asymptotic normality. One has to invest in theorems with assumptions which can be verified. Theoretical results are for time-continuous and time-discrete durations. Applications are for human and business demography, namely from a panel of claims data by a German health insurance company and from panel data of the German Statistical Office on enterprise foundation and closure. It is found that after a stroke, with time measured in years, the intensity of dementia onset increases from 0.02 to 0.07, with the standard error of the difference being 0.00093. It is found that the life expectancy of German enterprises is 10 years, with a standard error of 0.015. Both models are asymptotically normal. Events after the last wave are right censored and taken into account as usual.

C0247: **Testing truncation dependence and goodness-of fit for double-truncated durations**
Presenter: **Anne-Marie Toparkus**, Universitat Rostock, Germany
Co-authors: Rafael Weissbach

When analyzing left- or double-truncated durations, the log-likelihood can be derived from standard results for point processes, and the unobserved sample size can be profiled. Simplifying assumptions include the duration of interest being exponentially distributed or the age at truncation being independent of the duration. Both can be unappealing, so a nonparametric test for the first assumption is derived and a parametric test for the second. The second assumption is problematic when truncation results from data collection within a restricted time period, making the age at truncation equivalent to the date of birth, which conflicts with the progress of life expectancy. The dependence model uses the Gumbel-Barnett copula. Since the hypothesis parameter lies at the boundary, the test statistic's asymptotic distribution is a mixture of a two- and a one-dimensional normal distribution. For the first assumption, a bivariate Kolmogorov-Smirnov goodness-of-fit test is necessary for the additional attribute of age-at-truncation. Firstly, the truncated process' asymptotic behavior is analyzed when the true parameter is known under the null hypothesis. Replacing the unknown true parameter with its estimator alters the test statistic's distribution. The compactness of the truncation region allows for the computation of the test statistic using methods for the bivariate uniform case. Empirical examples include 55,000 lifetimes of German enterprises that ended between 2014 and 2016.

C0892: **Regression and prediction for competing risks with doubly truncated data**
Presenter: **Carla Moreira**, University of Minho, Portugal
Co-authors: Jacobo de Una-Alvarez

Regression and prediction for competing risks have been traditionally performed through a proportional hazards assumption. Such proportionality assumption can be established for the cause-specific hazards or transition intensities. In this objective, the approach is investigated when the target time is doubly truncated. More specifically, the applicability of the inverse probability weighting approach presented recently in literature in the competing risks setup is studied. The properties of the introduced procedures are investigated through simulations. Estimation, testing and prediction issues are considered. For illustration purposes, the proposed methods are applied to real data.

C0858: **Goodness-of-fit testing with survival data**
Presenter: **Jacobo de Una-Alvarez**, University of Vigo, Spain
Co-authors: Juan Carlos Escanciano

The aim is to present a new general strategy for goodness-of-fit testing with survival data. The setting is that of testing for a parametric family of distribution functions when the data deteriorates due to random censoring and/or random truncation. A key step is the characterization of the null hypothesis through a moment equation, which involves the estimation of the observable distribution under both the null and the alternative. An omnibus test based on a maximum mean discrepancy principle will be proposed, and its theoretical properties will be presented. The finite sample performance of the proposed test will be investigated through simulations. Illustrative real-data applications will be given.

CO156 Room S0.03 NEW ADVANCES ON COMPUTATIONALLY EFFICIENT STATISTICAL INFERENCE (VIRTUAL) Chair: Chong Jin

C1262: **Ensemble LDA via the modified Cholesky decomposition**
Presenter: **Zhenguo Gao**, Shanghai Jiao Tong University, China
Co-authors: Xinye Wang, Xiaoning Kang

A binary classification problem in the high-dimensional settings is studied via ensemble learning with each base classifier constructed from the linear discriminant analysis (LDA), and these base classifiers are integrated by the weighted voting. The precision matrix in the LDA rule is estimated by the modified Cholesky decomposition (MCD), which is able to provide a set of precision estimates by considering multiple variable orderings and, hence, yield a group of different LDA classifiers. Such available LDA classifiers are then integrated to improve the classification performance. The simulation and the application studies are conducted to demonstrate the merits of the proposed method.

C1452: **Improving the prediction of polygenic risk score: Overcoming challenges toward precision medicine**
Presenter: **Yiming Luo**, Columbia University, United States

Polygenic risk score (PRS) is a linear weighted combination of estimated effects from genetic variants and is a promising tool for personalized medicine. Advanced PRS methods have been developed to aggregate signals across the genome by using various penalization regression techniques. However, significant challenges remain in improving the performance of PRS across traits and populations. The purpose is to present the general principles and leading methods addressing three key areas: enhancing the transferability of PRS across populations, leveraging information from correlated traits, and incorporating rare genetic variants. By tackling these challenges, the aim is to improve the accuracy and applicability of PRS, furthering its potential as a valuable tool in precision medicine and genetic risk prediction.

C1645: **Using tissue-specific genetic variation in Mendelian randomization**
Presenter: **Chong Jin**, New Jersey Institute of Technology, United States
Co-authors: Ryan Liu, Qi Long

Studies have shown that intelligence is negatively correlated with Alzheimer's disease risk. However, the causality between tissue-mediated intelligence and Alzheimer's disease risk remains unclear. In addition, the tissue-related molecular pathways that mediate the effect of intelligence on Alzheimer's disease risk remain unclear. A two-sample multivariable Mendelian randomization study is conducted to analyze how tissue-specific gene expression mediates the effect of intelligence on Alzheimer's disease risk. Evidence that educational attainment independently affected Alzheimer's disease risk when analyzed with intelligence was not revealed. However, tissue-specific expression-mediated intelligence was associated with Alzheimer's disease risk independently from educational attainment with a log odds ratio of -1.59 (95% CI: -2.73, -0.443; p = 6.5710-3). The aim is to provide evidence supporting a causal association between tissue-specific gene expressions and the effect of intelligence on Alzheimer's Disease risk. Specifically, genes associated with intelligence located in brain tissues were found to be negatively correlated with Alzheimer's disease risk.

C1096: **Adaptive-TMLE for the average treatment effect based on randomized controlled trial augmented with real-world data**
Presenter: **Sky Qiu**, University of California, Berkeley, United States

The problem of estimating the average treatment effect (ATE) is considered when both randomized control trial (RCT) data and real-world data (RWD) are available. The ATE estimand is decomposed as the difference between a pooled-ATE estimand that integrates RCT and RWD and a bias estimand that captures the conditional effect of RCT enrollment on the outcome. An adaptive targeted minimum loss-based estimation (A-TMLE) framework is introduced to estimate them. The A-TMLE estimator is proven $\sqrt{n}$-consistent and asymptotically normal. Moreover, in finite sample, it achieves the super-efficiency one would obtain had one known the oracle model for the conditional effect of the RCT enrollment on the outcome. Consequently, the smaller the working model of the bias induced by the RWD is, the greater the estimator's efficiency, while the estimator will always be at least as efficient as an efficient estimator that uses the RCT data only. A-TMLE outperforms existing methods in simulations by having smaller mean-squared-error and 95% confidence intervals. A-TMLE could help utilize RWD to improve the efficiency of randomized trial results without biasing the estimates of intervention effects. This approach could allow for smaller, faster trials, decreasing the time until patients can receive effective treatments.

CO013 Room S0.11 ADVANCES IN COMPUTATIONAL NEUROSCIENCE Chair: Sharmistha Guha

C1689: **Deep kernel learning based Gaussian processes for Bayesian image regression analysis**
Presenter: **Jian Kang**, University of Michigan, United States

Regression models are widely used in neuroimaging studies to learn associations between clinical variables and image data. Gaussian process (GP) priors are common Bayesian nonparametric approaches to model unknown regression functions with complex spatial correlations in high-dimensional data. Existing GP methods depend on pre-specified parametric covariance kernel functions, which often have insufficient flexibility to capture data complexity, limited generalizability across study populations, and computational bottlenecks in large-scale datasets. We propose a scalable fully Bayesian kernel learning framework for GP priors with applications in various image regression models. Under kernel series expansion, we utilize the estimation power of deep neural networks (DNNs) to adaptively learn basis functions from data. We establish theoretical properties of posterior concentration rates in estimating regression and kernel functions. Through simulation studies, we show improved estimation and signal detection accuracy across different regression model settings. We illustrate the proposed method by analyzing multiple neuroimaging datasets in different medical studies.

C1688: **Joint Bayesian additive regression trees for inference of multiple non-linear dependence networks**
Presenter: **Christine Peterson**, The University of Texas MD Anderson Cancer Center, United States

Many diseases are heterogeneous, with subgroups of patients that have differing biological mechanisms. Estimating a single graph using the entire dataset may miss subtype-specific relations, while analyzing the data separately for each subgroup reduces statistical power to identify shared mechanisms. Importantly, the dependence relations among features may be non-linear. To address this challenge, we propose a hierarchical Bayesian model that encourages shared edge selection while allowing each group to have its own subtype-specific network. We formulate a flexible model based on the Bayesian additive regression tree framework to allow for non-linear dependencies within each subgroup, and encourage the joint selection of edges across groups through a hierarchical prior. The proposed method will be illustrated through both simulation studies and a real data application to protein-protein interaction networks across colorectal cancer subtypes.

C1576: **A Bayesian approach to simultaneous estimation of neural functional connectivity and sub-network structure**
Presenter: **Julia Fisher**, University of Arizona, United States
Co-authors: Edward Bedrick

Over the past couple of decades, a great deal of research has examined how distinct regions of the brain co-activate, typically using functional magnetic resonance imaging data gathered from subjects scanned while awake but at rest. A key observation of such studies has been that the functional network of the human brain can be partitioned into sets of regions (e.g., modules or sub-networks such as the well-studied default mode network) that are more closely related to each other than to regions outside the set. Most approaches to analyzing such data focus either on the estimation of region-to-region connectivity (often via partial or full correlation) or sub-network structure, but not both. Simultaneous estimation of the partial correlation between regions of interest and the partitioning of the network into modules is proposed via a Bayesian multivariate normal-Wishart model with an embedded Dirichlet process mixture model for clustering. The model on a range of simulated data is evaluated, and it is compared to other analytical approaches.

C1686: **Supervised modeling of heterogeneous networks**
Presenter: **Sharmistha Guha**, Texas A&M University, United States

The aim is to present a neuroimaging-driven study examining the relationship between functional connectivity across cognitive control domains and cognitive phenotypes, aiming to identify specific brain regions significantly associated with these phenotypes. We propose a generalized linear modeling framework incorporating multiple network responses and predictors, allowing for diverse interconnections between edges. Leveraging hierarchical Bayesian modeling, our approach estimates regression coefficients and identifies predictor-linked nodes with precise uncertainty quantification. Empirical investigations, including simulation studies and functional connectivity data analysis, demonstrate our framework's superior performance compared to competitors.

CO110 Room S0.12 MACHINE LEARNING METHODS IN EXTREMES Chair: Abdelaati Daouia

C0548: **Extremal random forests**
Presenter: **Nicola Gnecco**, University of Geneva, Switzerland
Co-authors: Sebastian Engelke, Edossa Merga Terefe

Classical methods for quantile regression fail in cases where the quantile of interest is extreme and only a few or no training data points exceed it. Asymptotic results from extreme value theory can be used to extrapolate beyond the range of the data, and several approaches exist that use linear regression, kernel methods or generalized additive models. Most of these methods break down if the predictor space has more than a few dimensions or if the regression function of extreme quantiles is complex. A method is proposed for extreme quantile regression that combines the flexibility of random forests with the theory of extrapolation. The extremal random forest (ERF) estimates the parameters of a generalized Pareto distribution, conditional on the predictor vector, by maximizing a local likelihood with weights extracted from a quantile random forest. The shape parameter is penalized in this likelihood to regularize its variability in the predictor space. Under the general domain of attraction conditions, consistency of the estimated parameters is shown in both the unpenalized and penalized case. Simulation studies show that the ERF outperforms both classical quantile regression methods and existing regression approaches from extreme value theory. The methodology is applied to extreme quantile prediction for U.S. wage data.

C0746: **An optimal index insurance framework for extreme losses**
Presenter: **Daniel Nkameni**, CREST ENSAE (Polytechnic Institute of Paris), France

The modern landscape of insurance is characterized by the emergence of new risks with destructive potential, challenging the feasibility of developing sustainable financial protection. Several administrations and regulators regularly highlight the potential uninsurability of certain risks, such as those associated with natural disasters in the context of climate change or cyber-attacks in the context of rapidly advancing artificial intelligence. To address these challenges, index-based insurance is frequently mentioned as a technical tool that may help provide coverage in these seemingly desperate situations. The construction of index insurance coverage against extreme losses is proposed, focusing on cases where these losses follow heavy-tail distributions. This index coverage is designed to activate above a certain threshold, while classical indemnity-based insurance is maintained below this threshold, leveraging the advantages of both types of coverage. The proposed methodology begins by adapting the usual utility of the wealth framework to the situation of heavy-tail losses. Subsequently, regression trees are employed to build the optimal index payout function through utility maximization. The final step involves determining the optimal threshold for the transition between the two types of coverage.

C0912: **Approximation principle for domain generalization**
Presenter: **Gloria Buritica**, AgroParisTech, France
Co-authors: Sebastian Engelke

In climate science studies, predicting the future impact of climate variables is crucial. In a climate change context, environmental variables like temperature and precipitation amounts increasingly reach extreme levels, highlighting the need for accurate predictions in these extreme scenarios.

Machine learning algorithms are popular for regression due to their strong predictive power on test points, which are well-represented during training. Their effectiveness comes from their ability to interpolate data without strict assumptions. However, these methods often fail to generalize well to underrepresented test points, struggling to predict extreme or rare events, as machine learning algorithms are not tailored to extrapolate. A median regression method is presented that improves predictions when covariates take extreme values. The approach is based on a regression extrapolation principle that models the response variable as covariates reach their highest levels. It works under minimal non-parametric restrictions, allowing satisfactory generalization results to be achieved.

C1242: **Bootstrapping block maxima estimators**
Presenter: **Torben Staud**, Ruhr University Bochum, Germany
Co-authors: Axel Buecher

Empirical means of disjoint block maxima calculated from a stationary time series are recognized to have a larger asymptotic variance than their counterparts based on sliding block maxima. The asymptotic variance formula for sliding block maxima involves the covariance of a certain bivariate Marshall-Olkin copula, and its estimation presents a complex problem. Standard plug-in methods are only feasible when explicit calculations are possible. As an alternative, reliance on bootstrap approximations is proposed. It is shown that naive block-bootstrap approaches are inconsistent, and a consistent alternative is provided based on resampling circular block maxima. These are obtained by calculating sliding maxima on small blocks of blocks (for instance, twice the block size r) after repeating the first $r-1$ observations in each such block of blocks. Additionally, it is demonstrated that the circular block maxima method shares the same asymptotic variance as the sliding block maxima method. The superiority of circular block maxima-based bootstrap methods over disjoint blocks-based methods is verified through large-scale simulation studies, and its usefulness is illustrated in a case study concerning precipitation at a fixed location.

CO125 Room S0.13 RECENT ADVANCES IN BIOSTATISTICS Chair: Yuedong Wang

C0265: **An alternative measure for quantifying the heterogeneity in meta-analysis**
Presenter: **Tiejun Tong**, Hong Kong Baptist University, Hong Kong

Quantifying the heterogeneity is an important issue in meta-analysis, and among the existing measures, the I^2 statistic is most commonly used. A simple example first illustrates that the I^2 statistic is heavily dependent on the study sample sizes, mainly because it is used to quantify the heterogeneity between the observed effect sizes. To reduce the influence of sample sizes, an alternative measure is introduced that aims to directly measure the heterogeneity between the study populations involved in the meta-analysis. A new estimator is further proposed, namely the I_A^2 statistic, to estimate the newly defined measure of heterogeneity. For practical implementation, the exact formulas of the I_A^2 statistic are also derived under two common scenarios with the effect size as the mean difference (MD) or the standardized mean difference (SMD). Simulations and real data analysis demonstrate that the I_A^2 statistic provides an asymptotically unbiased estimator for the absolute heterogeneity between the study populations, and it is also independent of the study sample sizes as expected.

C1008: **Estimating causal effects with proximal inference methods in single-cell CRISPR screens**
Presenter: **Catherine Wang**, Carnegie Mellon University, United States
Co-authors: Kathryn Roeder, Larry Wasserman

Single-cell CRISPR screen experiments allow for the tests and estimation of causal effects by perturbing a genomic region of interest and measuring changes in gene expression; however, insufficient adjustment for confounding may result in bias, producing large false positive rates. Typically, when the confounding variables are unmeasured, casual estimands cannot be identified and estimated, but the proximal setup allows for causal inference by taking advantage of two types of proxy variables: treatment-inducing confounding proxies and outcome-inducing confounding proxies. In the single-cell CRISPR setting, scientists measure many CRISPR perturbations and genes' expressions that may act as such proxies. A methodology for analyzing causal effects between many perturbation-gene pairs using proximal inference is proposed. Estimation procedures are explored, including two-stage least squares approaches, generalized method of moments approaches, and a plug-in and one-step corrected approach. These are designed to take advantage of the wide selection of proxies, each of which may be individually weak. Finally, the assumptions required for proximal inference are discussed and evaluated, and results are shown in simulated and real-world single-cell CRISPR screens.

C1352: **Penalized deep partially linear Cox models**
Presenter: **Yi Li**, University of Michigan, United States

A novel penalized deep partially linear Cox model (Penalized DPLC) is proposed, which incorporates the SCAD penalty to select important texture features and employs a deep neural network to estimate the nonparametric component of the model. The convergence and asymptotic properties of the estimator are proven and are compared to other methods through extensive simulation studies, evaluating its performance in risk prediction and feature selection. The proposed method is applied to the NLST study dataset to uncover the effects of key clinical and imaging risk factors on patients' survival. Findings provide valuable insights into the relationship between these factors and survival outcomes.

C1456: **Time-to-event analysis with unknown time origins via longitudinal biomarker registration**
Presenter: **Wensheng Guo**, University of Pennsylvania, United States

In observational studies, the time origin of interest for time-to-event analysis is often unknown, such as the time of disease onset. Using the study entry as time zero often leads to misleading results. Existing approaches to estimating the time origins are commonly built on extrapolating a parametric longitudinal model, which relies on rigid assumptions that can lead to biased inferences. A flexible semiparametric curve registration model is introduced. It assumes the longitudinal trajectories follow a flexible common shape function with a person-specific disease progression pattern characterized by a random curve registration function, which is further used to model the unknown time origin as a random start time. This random time is used as a link to jointly model the longitudinal and survival data where the unknown time origins are integrated into the joint likelihood function, which facilitates unbiased and consistent estimation. Simulation studies and two real data applications demonstrate the effectiveness of this new approach.

CO178 Room Safra Lec. Theatre METHODS FOR ANALYZING STRUCTURED DATA Chair: Israel Almodovar Rivera

C0257: **Exploratory factor analysis of data on a sphere**
Presenter: **Fan Dai**, Michigan Technological University, United States
Co-authors: Karin Dorman, Somak Dutta, Ranjan Maitra

Data on high-dimensional spheres arise frequently in many disciplines either naturally or as a consequence of preliminary processing and can have intricate dependence structures that need to be understood. Exploratory factor analysis of the projected normal distribution is developed to explain the variability in such data using a few easily interpreted latent factors. The methodology provides maximum likelihood estimates through a novel fast alternating expectation profile conditional maximization algorithm. Outputs on simulation experiments in a wide range of settings are uniformly excellent. The method gives interpretable and insightful results when applied to tweets with the "me too" hashtag in 2018, to time-course functional magnetic resonance images of the average pre-teen brain at rest, to characterize handwritten digits, and to gene expression data from cancerous cells in the cancer genome atlas.

C0268: **On data analysis pipelines and modular Bayesian modeling**
Presenter: **Abel Rodriguez**, University of Washington, United States

The most common approach to implementing data analysis pipelines involves obtaining point estimates from the upstream modules and then treating these as known quantities when working with the downstream ones. This approach is straightforward, but it is likely to underestimate the overall uncertainty associated with any final estimates. An alternative approach involves estimating parameters from the modules jointly using a Bayesian hierarchical model, which has the advantage of propagating upstream uncertainty into the downstream estimates. However, when modules are misspecified, such a joint model can behave in unexpected ways. Furthermore, hierarchical models require the development of ad-hoc computational implementations that can be laborious and computationally expensive. Cut inference modifies the posterior distribution to prevent information flow between certain parameters and provides a third alternative for statistical inference in data analysis pipelines. A unified framework is presented that encompasses two steps, cut and joint inference, in the context of data analysis pipelines with two modules. It also uses two examples to illustrate the tradeoffs associated with these approaches. It is shown that cut inference provides both some level of robustness and ease of implementation for data analysis pipelines at a lower cost in terms of statistical inference.

C0967: **New methods to genotype allopolyploids**

Presenter: **Karin Dorman**, Iowa State University, United States

Many economically important plants are allopolyploids, carrying multiple subgenomes or homoeologous copies of each chromosome, typically derived from two or more ancient diploid ancestors that combined to form the allopolyploid. Recent sequencing efforts have yielded reference subgenomes, thus inviting new strategies for genotyping allopolyploids from next-generation sequencing data. A genotyping method was recently introduced that models the hierarchical structure on the read alignments, showing that it could achieve better performance than algorithmic methods that split the reads by sequence similarity to the subgenomes. The method is extended to also handle the structural dependence among sites along the chromosomes. The result is nested layers of hidden information with inference by a modified EM algorithm. The method for synthetic allotetraploids is tested (real data, truth-known) and the new method is shown to improve over existing allopolyploid genotyping methods.

C0971: **Gauge reproducibility and repeatability for matrix-variate data with application to forensic fracture surface-matching**

Presenter: **Ranjan Maitra**, Iowa State University, United States

Co-authors: Carlos Llosa-Vite

Three-dimensional (3D) microscopy can be used to analyze the unique microscopic patterns present in fractured surfaces and, therefore, help forensic match-analysts reduce the subjectivity involved in comparative microscopy. Yet, the repeatability and reproducibility of these 3D microscopy-generated features have seldom been studied, and little is known about the effects that the microscope operator or sample alignment has on the measurement system. To study the quality of the measurement system, three inexperienced microscope operators are trained, and six overlapping images are repeatably obtained from a set of 10 steel rods that were generated and fractured under controlled conditions. A novel gauge R&R model is developed for matrix-variate data and proposed an expectation-maximization algorithm for maximum likelihood estimation to quantify the multiple sources of variability affecting the obtained features. After the third imaging repetition, each operator used a fixture to align the two fractures during the imaging process. While all the matches and non-matches were classified correctly regardless of the imaging fixture, gauge R&R showed that the fixture helped us improve the measurement system by keeping the within-operator as the smallest source of variability. The importance of assessing 3D microscopy measurement systems is shown with tools such as gauge R&R, as it can help improve the measurement system.

CO026 Room BH (S) 1.01 Lec. Theatre 1 TOPICS IN TIME SERIES ECONOMETRICS Chair: Johan Lyhagen

C0665: **Statistical modelling of variables on the unit interval (the closed interval between zero and one)**

Presenter: **Jonas Andersson**, Norwegian School of Economics, Norway

Variables representing fractions sometimes have values that are exactly zero or exactly one. This prohibits the direct use of some probability distribution, e.g., the beta distribution, as an appropriate representation. I will discuss the modelling of such variables. This is exemplified by using the experimental data collected by a prior study. Some aspects of the drivers of the declaration rate are shown by using a zero-one inflated beta regression model. It turns out that some of the drivers of the declaration rate affect the three parts of the declaration rate distribution, the zero declarers, the full declarers and the intermediate declarers, differently. It is found that the effect of taxpayers' monitoring, i.e., their knowledge about other taxpayers' evasion, increases the probability of declaring zero. Among the individuals declaring a part of their income, the effect is significantly positive; they declare more. Another result is that, for the average experiment participant, the probability of fully declaring or declaring nothing of income increases as the experiment progresses. Furthermore, time dependence in models on [0,1] is discussed.

C0753: **An extension of likelihood ratio based tests for evaluating the interval forecast**

Presenter: **Yushu Li**, University of Bergen, Norway

Evaluation of the forecast is a fundamental concern in time series forecasting, and the purpose of the evaluation is to check if the ex-post realization aligns with the ex-ante forecast. The likelihood ratio-based tests for evaluating the interval forecast are widely implemented to evaluate the interval forecast, especially in finance risk modelling. For example, value-of-risk (VaR) estimates are an application of one-sided interval forecasting, and the Christoffersen test can be used to test the validity of VaR forecasts in, for example, GARCH models. The likelihood ratio test framework has also been extended to evaluate the density forecast. Although the prior study mentioned their univariate test technique could be extended to multivariate straightforward, no previous study has been done to extend the test to bivariate case time series cases where the forecast area will be a two-dimensional region instead of a one-dimensional interval. The aim is to fill the gap with a concrete extended test framework, investigation of the size and power of the test, and empirical implementations.

C0933: **Beta regression: Shrinkage-Liu type estimator with application**

Presenter: **Peter Karlsson**, Linneaus University, Sweden

Co-authors: Stanislas Muhinyuza, Maziar Sahamkhadam

Beta regression is widely used for modeling outcome variables bounded within the open interval from zero to one. However, when non-orthogonal regressors are present, the performance of the maximum likelihood estimator deteriorates, and a common solution to address this issue is to use a shrinkage estimator. Furthermore, sometimes, there is prior information about the parameters in the Beta regression model, such that the parameter vector is suspected to belong to a linear subspace. The restricted linear regression model with multicollinearity is common in practice, and shrinkage estimators of the model coefficients have been considered in the literature. However, the literature on shrinkage estimators for Beta regression models is limited under such settings. A two-parameter Liu linear shrinkage estimator is introduced, tailored for estimating the vector of parameters in a Beta regression model with a fixed dispersion parameter under the assumption of linear restrictions on the parameter vector. This estimator is particularly applicable in various practical scenarios where the level of correlation among the regressors varies, and the coefficient vector is suspected to belong to a linear subspace. Furthermore, the necessary and sufficient conditions for establishing the superiority of the new estimator over both one-parameter Liu estimators and two-parameter Stein-type estimators are derived. Finally, an empirical application is presented.

C0943: **Robust LM-type testing in multivariate contaminated time series**

Presenter: **Yukai Yang**, Uppsala University, Sweden

The challenging problem of performing robust Lagrange Multiplier (LM) tests is addressed in multivariate contaminated time series, where unobservable additive and innovative outliers present significant difficulties. The complexities of detecting and managing outliers within time series models are highlighted. A novel algorithm is introduced, designed to be compatible with any subsample-based LM test, thereby creating a robust version of the LM test. The sufficient conditions required for the algorithm and the modified test are derived to achieve robustness and consistency in the presence of outliers, ensuring reliable performance in large samples. Simulation studies showcase its superior performance and robustness in finite samples. Two real-world applications further demonstrate the practical effectiveness of the algorithm and the resulting robust LM test.

CO064 Room BH (SE) 1.01 ADVANCES IN HIGH-DIMENSIONAL BAYESIAN INFERENCE Chair: David Rossell

C0206: **Efficient Gibbs sampling for latent space models**

Presenter: **Roberto Casarin**, University Ca' Foscari of Venice, Italy
Co-authors: Antonio Peruzzi

Latent space (LS) network models project the nodes of a network on a d-dimensional latent space to achieve dimensionality reduction of the network while preserving its relevant features. Inference is often carried out within a Markov chain Monte Carlo (MCMC) framework. Nonetheless, it is well-known that the computational time for this set of models increases quadratically with the number of nodes. The purpose is to build on the random-scan (RS) approach to propose an MCMC strategy that alleviates the computational burden for LS models while maintaining the benefits of a general-purpose technique. The novel sampling strategy effectively reduces the computational cost by a factor without severe consequences on the MCMC draws. Some convergence properties of the MCMC procedure are provided, and it is shown via simulation that this RS approach performs better than the standard RS in terms of mixing. Finally, the sampler is applied to a multi-layer temporal LS model, and it is shown how the adaptive strategy may be beneficial to empirical applications.

C0600: **Bayesian transfer learning with multiple auxiliary datasets**

Presenter: **Donatello Telesca**, UCLA, United States

Transfer learning is considered in the context of high dimensional linear and generalized linear models. While Bayesian inferential methods are naturally suited for TL, some care is needed in the construction of sparsity-inducing priors to mitigate the effects of possible negative transfers. When multiple auxiliary datasets are available to inform a target task, we show how the combination of regularization priors with standard Bayesian model selection can prove effective in the identification of an informative auxiliary set while accounting for uncertainty in the selection of important covariates. A direct parametrization through sparse contrasts allows for fine-tuning of the level of borrowing and desired levels of model compatibility. Finally, parallels are discussed with related techniques in Bayesian meta-analysis and several classes of power/power-like priors.

C0809: **The importance of the tails in high dimensional macroeconomic forecasting**

Presenter: **Anna Simoni**, GENES - CREST, France
Co-authors: Matteo Mogliani, Luca Rossini

Forecasting beyond the central tendency of future values of macroeconomic and finance series is important for policymakers, as it allows for quantifying the likelihood of tail (upside or downside) risks. Nowadays, policymakers have available rich datasets that potentially contain valuable information for predicting tail events but that are difficult to exploit due to the large dimension. A large-dimensional quantile regression model is proposed to forecast the tails of the conditional distribution of future values of the target variable, and at the same time, the large dimension is dealt with by exploiting a bi-level sparse group structure as well as a regularization scheme. A Bayesian procedure is proposed based on the asymmetric Laplace distribution, which is shown to have optimality properties. While the group structure is assumed to be known, the sparsity structure is not, and it is shown that the approach learns adaptively about which groups and which variables are active. Interestingly, the sparsity structure is quantile-dependent. Frequentist asymptotic properties of the procedure are studied. Finite sample properties are illustrated through Monte Carlo experiments. Finally, the performance of the procedure is shown with real macroeconomic data.

C0891: **External information for high-dimensional variable selection**

Presenter: **Paul Rognon-Vael**, U. Pompeu Fabra - U. Politecnica de Catalunya, Spain
Co-authors: David Rossell

In many modern applications, the sample size is often insufficient to estimate the parameters of a model reliably. This limitation has motivated the development of high-dimensional techniques that frequently rely on very strict assumptions on sparsity and signal size. Yet, often, external information is available and can be leveraged to enhance parameter estimation. This concept is investigated in the context of variable selection in basic linear regression scenarios. It is shown that incorporating external information allows pushing the theoretical limits under which consistent variable selection is possible. A concrete model selection procedure based on Bayesian principles is proposed, which realizes those benefits and outperforms standard penalization.

CO121 Room BH (SE) 1.02 MODERN CHALLENGES IN BAYESIAN NONPARAMETRIC INFERENCE Chair: Federico Camerlenghi

C0450: **Bayesian adaptive Tucker decompositions for tensor factorization**

Presenter: **Federica Stolf**, University of Padova, Italy
Co-authors: Antonio Canale

A novel Bayesian method is presented for low-rank tensor decomposition of multiway data with missing observations. An adaptive Tucker decomposition model (AdaTuck) is introduced that automatically infers the latent multi-rank in a principled manner using an increasing shrinkage prior. The proposed approach introduces local sparsity in the core tensor through a shrinkage prior, inducing rich and, at the same time, parsimonious dependency structures. Posterior inference proceeds via an efficient adaptive Gibbs sampler, allowing for straightforward missing data imputation and making it well-suited for high-dimensional datasets. Simulation studies and applications to complex spatiotemporal data provide compelling support for the new modeling class relative to existing tensor factorization methods.

C0468: **Loss-based prior for tree topologies in BART models**

Presenter: **Fabrizio Leisen**, Kings College London, United Kingdom
Co-authors: Cristiano Villa, Kevin Wilson, Francesco Serafini

The purpose is to present a novel prior for tree topology within Bayesian additive regression trees (BART) models. This approach quantifies the hypothetical loss in information and the loss due to the complexity associated with choosing the wrong tree structure. The resulting prior distribution is compellingly geared toward sparsity, a critical feature considering BART models' tendency to overfit. The method incorporates prior knowledge into the distribution via two parameters that govern the tree's depth and balance between its left and right branches. Additionally, a default calibration is proposed for these parameters, offering an objective version of the prior. The method's efficacy is demonstrated on both simulated and real datasets.

C0476: **Random signed measures**

Presenter: **Riccardo Passeggeri**, Imperial College London, United Kingdom

Point processes and, more generally, random measures are ubiquitous in modern statistics. However, they can only take positive values, which is a severe limitation in many situations. Random signed measures are introduced, also known as real-valued random measures. In particular, an existence result is provided for random signed measures, allowing the obtainment of a canonical definition for them, and a 70-year-old open problem is solved. Further, a representation of completely random signed measures (CRSMs) is provided, which extends the celebrated Kingman's representation of completely random measures (CRMs) to the real-valued case. Specific classes of random signed measures are then presented, including the Skellam point process, which plays the role of the Poisson point process in the real-valued case, and the Gaussian random measure. These measures are used to construct a Bayesian nonparametric model and a sparse random signed graph model and to explore mean function estimation in nonparametric regression.

C0785: **Latent random partition model: An application to childhood co-morbidity**
Presenter: **Maria De Iorio**, National University of Singapore, Singapore

Asthma, hypertension and obesity are three of the most common chronic diseases worldwide, with known presence of comorbid pathophysiological mechanisms. Such data are collected from different sources and are usually analysed separately, neglecting the shared information among subjects, underlining the need for a more comprehensive statistical approach. A novel Bayesian nonparametric model is developed for the joint analysis of biomarkers of different types related to obesity (longitudinal data), history of asthma (panel count data) and symptoms of hypertension (multi-state process). In particular, the random partitions of the subjects are modelled in each dataset independently and conditionally on an underlying partition structure. The proposed strategy allows for the sharing of information among the clustering structures within the different datasets, thus providing more robust inference. Random partitions of different datasets are marginally dependent, with dependence learnt from the data. The model allows for the inclusion of mixed-type covariates, aiding the identification of risk factors affecting the evolution of the diseases. A tailored MCMC algorithm is developed. The model is demonstrated in an application from the Singaporean birth cohort GUSTO.

CO168 Room BH (SE) 1.05 Recent developments in financial modelling and forecasting Chair: Spyros Vrontos

C1047: **Convergence in academic productivity**
Presenter: **Theologos Pantelidis**, University of Macedonia, Greece
Co-authors: Maria Karantali, Theodore Panagiotidis

Academic productivity is a key indicator for researchers, educational institutions, and society as a whole, and it has a direct impact on the socio-economic growth of a country through the appropriation of knowledge. The existence of (club) convergence in academic productivity among various groups of countries is investigated by means of the methodology developed in past studies. More specifically, the focus is on (i) the OECD countries and (ii) the G20 countries, and the dataset covers a period of more than 25 years. Convergence in publications (per 1000 people) to measure the output level and convergence in citations (either per 1000 people or per publication) is examined to measure the impact of the research output. In a few cases, full convergence is supported, while in all other cases, convergence clubs are identified. Various robustness tests evaluate the sensitivity of the main findings.

C1071: **Pairs trading using machine learning models**
Presenter: **Spyros Vrontos**, University of Essex, United Kingdom

The application of machine learning models to pairs trading is explored, a popular market-neutral trading strategy that involves taking simultaneous long and short positions in two correlated securities. The approach leverages advanced machine learning techniques to enhance the selection and timing of trades. By employing machine learning models, the aim is to improve the prediction accuracy of price movements and the overall profitability of the strategy. The performance of these models is evaluated using historical market data, comparing their effectiveness against conventional pairs trading methods.

C1072: **Enhanced covariance matrix estimators in portfolio management: Comparing parametric and nonparametric approaches**
Presenter: **Sarah Alsaed**, University of Essex, United Kingdom
Co-authors: Spyridon Vrontos

Traditional mean-variance optimizers that rely on raw estimates of the covariance matrix tend to be unstable due to the substantial amount of noise in the sample covariance matrix. The aim is to refine and enhance the covariance matrix used in portfolio optimization problems. Shrinkage techniques, applications of random matrix theory, and a variety of hybrid approaches are examined. In addition, the parametric DCC GARCH model is examined to estimate dynamic conditional correlations. A comparative analysis using a rolling window approach is employed to test the efficiency of each covariance matrix estimator in achieving robustness and better performance metrics. A plethora of investment strategies is tested and evaluated.

CO008 Room BH (SE) 1.06 Theory and implementation of statistics for stochastic processes Chair: Hiroki Masuda

C0344: **Parameter estimation for linear parabolic SPDEs in two space dimensions based on high frequency spatiotemporal data**
Presenter: **Masayuki Uchida**, Osaka University, Japan
Co-authors: Yozo Tonaki, Yusuke Kaino

The problem of estimating unknown coefficient parameters of linear parabolic second-order stochastic partial differential equations (SPDEs) is treated in two space dimensions driven by Q-Wiener processes using high-frequency spatiotemporal data. Recent studies investigated minimum contrast estimators (MCEs) for unknown coefficient parameters of a linear parabolic second-order SPDE in one space dimension driven by a cylindrical Wiener process based on high-frequency spatiotemporal data. The methodology of an existing study is applied to the SPDEs in two space dimensions, and MCEs of the coefficient parameters are introduced to the SPDEs in two space dimensions using temporal and spatial increments. The coordinate processes of the SPDEs are then approximated using the MCEs. Furthermore, utilizing the approximated coordinate processes, parametric adaptive estimators are proposed for the rest of the unknown parameters of the SPDEs. Numerical simulations of the proposed estimators are also provided.

C0364: **Locally stable approximation of SDE: Numerical algorithms in yuimaStable**
Presenter: **Lorenzo Mercuri**, University of Milan, Italy
Co-authors: Hiroki Masuda

The purpose is to analyze the numerical aspects related to the estimation of an ergodic Markovian stochastic differential equation (SDE) driven by a locally stable Levy process. The likelihood function constructed on the small-time stable approximation of the transition density (QSMLE) leads to an estimator with appropriate asymptotic behavior. To the best of knowledge, the joint estimation of drift, jump coefficients, and stability index has only been considered in recent literature. However, in previous works, the problem of filtering noise from the data has not been discussed from either a theoretical or numerical point of view. A two-step procedure is considered. First, the drift, the jump coefficient, and the stability parameters are estimated jointly. Consequently, the filtered increments are reconstructed with the possibility of estimating the Levy measure parameters for the underlying noise. To this aim, the transition stable-density is numerically evaluated in QSMLE using different Gauss-like quadrature methods. Finally, the new classes and methods implemented in yuimaStable for the estimation of these SDEs are discussed with simulated and real high-frequency data.

C0420: **Spectral calibration of time-inhomogeneous exponential Levy models**
Presenter: **Jakob Soehl**, Delft University of Technology, Netherlands
Co-authors: Loek Koorevaar, Stan Tendijck

Empirical evidence shows that calibrating exponential Levy models by options with different maturities leads to conflicting information. In other words, the stationarity implicitly assumed in the exponential Levy model is not satisfied. An identifiable time-inhomogeneous Levy model is proposed that does not assume stationarity and that can integrate option prices from different maturities and different strike prices without leading to conflicting information. In the time-inhomogeneous Levy model, the convergence rates are derived, and confidence intervals are shown for the estimators of the volatility, the drift, the intensity and the Levy density. Previously, confidence intervals have been constructed for time-homogeneous Levy models in an idealized Gaussian white noise model. In the idealized Gaussian white noise model, it is assumed that the observations are Gaussian and given continuously across the strike prices. This simplifies the analysis significantly. The confidence intervals are constructed in a discrete observation setting for time-inhomogeneous Levy models, and the only assumption on the errors is that they are sub-Gaussian. In particular, all bounded errors with arbitrary distributions are covered. Additional results on the convergence rates extend existing results from time-homogeneous to time-inhomogeneous Levy models.

C0469: **A note on uniform confidence bands for spot volatility**
Presenter: **Yuta Koike**, University of Tokyo, Japan

Using the Gaussian approximation theory for maximum type statistics (the so-called CCK theory), uniform confidence bands are constructed for the spot volatility of a discretely observed Ito semi-martingale in a simplified setting. The recent developments in the CCK theory improve the accuracy of the existing result, so this point is particularly demonstrated. The practical implementation is also discussed.

CO372 Room BH (S) 2.01 FINANCIAL ECONOMETRICS AND MACHINE LEARNING Chair: Xiaohan Xue

C1119: **Deep learning enhanced financial time series forecasting**
Presenter: **Chao Wang**, The University of Sydney, Australia
Co-authors: Minh-Ngoc Tran, Richard Gerlach, Robert Kohn

We investigate how to tailor and integrate the deep learning techniques into the conventional financial econometric models for financial risk forecasting. First, a long short-term memory enhanced realized conditional heteroskedasticity model is developed, to explore the full impact of high frequency data based realized volatility on volatility modeling and forecasting via capturing the nonlinear and long-term effects. Further, extending the heterogeneous autoregressive model, a framework known for efficiently capturing long memory in realized measures, a long-memory and non-linear realized volatility model class is proposed for direct Value-at-Risk forecasting by integrating the Recurrent Neural Network. Bayesian inference with Sequential Monte Carlo is employed for model estimation and sequential prediction in the proposed frameworks. Comprehensive empirical study using 31 indices from 2000 to 2022 is conducted. The results demonstrate that our proposed framework achieves superior out-of-sample risk forecasting performance compared to the benchmark models.

C1196: **GNN based social medium analysis in stock prediction**
Presenter: **Jin Zheng**, University of Bristol, United Kingdom
Co-authors: Pengju Zhang, Richard Harris

The purpose is to introduce a novel approach that utilises graph neural networks (GNNs) for sentiment analysis to enhance stock market predictions. By integrating social media sentiment analysis with traditional financial indicators, a comprehensive model that accurately captures market sentiment dynamics is developed. The methodology includes data collection from social media platforms, sentiment extraction using GNN, and prediction using advanced time-series models such as LSTM, CNN and Transformer. The model is evaluated on several stock datasets, and improvements are demonstrated in prediction accuracy and trading performance compared to traditional models. These results underscore the value of merging sentiment analysis with time-series techniques for financial market prediction. Furthermore, the use of the prediction model to improve trading strategy performance is discussed by comparing different trading strategies based on the daily return and Sharpe ratio. The contribution to the field is by providing insights into the role of sentiment in financial markets and by advancing the capabilities of predictive models through the integration of GNN-based sentiment analysis.

C1488: **The likelihood ratio test for changes in high-dimensional idiosyncratic network**
Presenter: **Xiaohan Xue**, University of Bath, United Kingdom

A novel framework is introduced, employing the likelihood ratio test to detect change points in high-dimensional networks. The framework introduces a sophisticated factor model for handling common variations in stock returns, employing tools like Graphical Lasso for estimating sparse precision matrices essential in high-dimensional settings. The simulations demonstrate the framework's effectiveness in controlling type I error rates and its power to detect true positives under various scenarios. Applying the methodology to the daily returns of S&P 500 constituents illustrates practical implications. The results identify multiple structural breaks coinciding with major economic events, underscoring the test's potential in real-world scenarios where detecting such breaks can significantly impact investment strategies and risk assessment.

C1646: **Backtesting expectile: Disentangling unconditional coverage and independence properties**
Presenter: **Xiaochun Meng**, University of Bath, United Kingdom
Co-authors: Yang Lu, Melina Mailhot, Jesus Armando de Ita Solis

Under current regulations, financial institutions are required to estimate the daily value-at-risk (VaR) or expected shortfall (ES) of their trading positions. Despite being widely studied and adopted, both risk measures have theoretical limitations: VaR is not coherent, and ES is not elicitable. Expectile, the only law-invariant risk measure that is both coherent and elicitable, has gained considerable interest in both risk management and statistics recently. However, the backtesting of expectile has not yet received adequate attention, and existing backtests tend to suffer from size distortion or low test power. The aim is to propose novel unconditional and conditional backtests for expectile. Simulation studies show that the proposed tests exhibit promising finite sample size performance. In addition, in an empirical study, the proposed tests are applied to S&P data.

CO385 Room BH (S) 2.02 HETEROGENEITY IN PANEL DATA MODELS Chair: Jiaying Gu

C1372: **Inference on union bounds**
Presenter: **Xinyue Bei**, The University of Texas at Austin, United States

A union bound is a union of multiple bounds. Union bounds occur in a wide variety of empirical settings, such as difference-in-differences, regression discontinuity design, bunching, and misspecification analysis. A confidence interval is proposed for these kinds of bounds based on modified conditional inference. It is improved upon existing methods in a large set of data-generating processes. The new procedures give statistically significant results, unlike the pre-existing alternatives in the empirical applications.

C1388: **Predicting unobserved individual-level causal effects**
Presenter: **Christophe Gaillac**, University of Geneva, Switzerland

Measuring accurately heterogeneous effects is key for the design of efficient public policies. The focus is on predicting unobserved individual-level causal effects in linear random coefficients models, conditional on all the available data. In the application, these "posterior effects" are considered the average effects of teachers' knowledge on their students' performance, conditional on both variables. Two nonparametric strategies are derived for recovering these posterior effects, assuming independence between the effects and the covariates. The first strategy recovers the distribution of the random coefficients by a minimum distance approach and then obtains the posterior effects from this distribution. The corresponding estimator can be computed using an optimal transport algorithm. The second approach, which is valid only for continuous regressors, directly expresses the posterior effects as a function of the data. The corresponding estimator is rate optimal. I discuss several extensions, in particular, the relaxation of the independence condition. Finally, the application reveals large heterogeneity in the effect of teachers' knowledge, suggesting that the cost-effectiveness of their training could be substantially improved.

C1471: **C(alpha) test for number of components in finite mixture models**
Presenter: **Junfan Tao**, Kyoto University, Japan
Co-authors: Jiaying Gu, Stanislav Volgushev

Finite mixture models are widely used in empirical applications, yet determining the number of components remains a challenging task. The aim is to propose a C(alpha) test for inference on the number of components. It is shown that it is asymptotically equivalent to the EM test but does not require tuning parameters and often yields better power performance in finite samples. Both the C(alpha) and EM tests have limiting distributions that are mixtures of chi-squares, though the weights typically require simulation. To address this, the most stringent somewhere most powerful test (MSSMP) is also proposed, as originally introduced in a past study. This test benefits from a pivotal limiting distribution. Simulation studies show that the MSSMP test possesses comparable performance to the C(alpha) test. An empirical example to demonstrate the application of these tests is concluded with.

C1536: **Identification of dynamic panel logit models with fixed effects**
Presenter: **Jiaying Gu**, University of Toronto, Canada

It is shown that the identification problem for a class of dynamic panel logit models with fixed effects has a connection to the truncated moment problem in mathematics. This connection is used to show that the sharp identified set of structural parameters is characterized by a set of moment equality and inequality conditions. This result provides sharp bounds in models where moment equality conditions do not exist or do not point at identifying the parameters. It is also shown that the sharp identified set of the non-parametric latent distribution of the fixed effects is characterized by a vector of its generalized moments and that the number of moments grows linearly in T. This final result point identifies, or sharply bound, specific classes of functionals, without solving an optimization problem with respect to the latent distribution. The identification result is illustrated with several examples and an empirical application for modelling children's respiratory conditions.

CO175 Room BH (S) 2.03 QUANTITATIVE CLIMATE ANALYSIS Chair: Jesus Gonzalo

C0185: **Global and regional long-term climate forecasts: A heterogeneous future**
Presenter: **Lola Gadea**, University of Zaragoza, Spain
Co-authors: Jesus Gonzalo

Climate is a long-term issue; therefore, climate forecasts should be long-run. Such forecasts are crucial for designing the mitigation policies required to fulfil one of the main objectives of the Paris Climate Agreement (PCA) and design adaptation policies that mitigate the adverse effects of climate change. They also serve as an indispensable instrument to assess climate risks and successfully steer the green transition. The aim is to propose a simple method to produce long-term temperature density forecasts from observational data using the realized quantile methodology introduced in an existing study, where unconditional quantiles are converted into time series objects. They complement the projections obtained by physical climate models, mainly focused on the mean temperature. These averages usually conceal wide spatial disparities, which, among other distributional characteristics, are captured by our density forecast. The proposed method consists of running an out-of-sample forecast model competition and combining the estimates of the resulting Pareto-superior models to eliminate the forecast model dependency. Furthermore, the approach considers climate change a non-uniform phenomenon; therefore, analyzing it from a regional perspective is crucial, offering different predictions for a heterogeneous future.

C0199: **An unconditional-quantile vector error correction model to analyze climate heterogeneity**
Presenter: **Andrey Ramos**, Carlos III University of Madrid, Spain

A novel time-series quantitative methodology is introduced to analyze heterogeneity in the temperature distribution and its association with climate forcings. The approach employs a vector autoregressive model (VAR) for a range of unconditional distributional characteristics of temperature mean and quantiles alongside the radiative forcing of greenhouse gases (GHGs), including carbon dioxide (CO2), methane (CH4), and nitrous oxide (N2O). The empirical analysis is conducted across three geographical scales: the Globe, the North Hemisphere, and Europe, utilizing station-level data from the Climatic Research Unit (CRU) during the period 1880-2021. At these scales, the temperature unconditional quantiles represent temperature at different latitudes, and the results are comparable to the predictions of one-dimensional (1D) energy balance models (EBMs). However, the methodology can be adapted to more general situations with limited spatial variation, assuming there is sufficient cross-sectional or higher frequency data to derive the unconditional distributional characteristics of temperature. The proposed methodology serves various purposes, including i) Estimation of physical parameters like climate sensitivity, ii) Forecasting, iii) Identification of structural shocks and impulse-response analysis, and iv) Predictions of temperature conditional on hypothetical emissions/concentrations scenarios.

C0281: **Testing extreme warming and geographical heterogeneity**
Presenter: **Jose Olmo**, Universidad de Zaragoza, Spain
Co-authors: Jesus Gonzalo, Lola Gadea

Extreme weather events represent a critical climate risk that presents a global challenge. Understanding the heterogeneity in worldwide temperatures is vital for predicting future climate change dynamics and guiding policy responses. Both issues are addressed by proposing analytical methods to study the dynamics of extreme temperatures and their geographical heterogeneity. As a secondary focus, the finite-sample performance of Hill-type estimators of tail decay for location-scale models is examined. It is found that estimators obtained from standardized order statistics demonstrate significantly better performance for heavy-tailed distributions under substantial location effects. Applying the proposed methodology, the presence of trends in the tail decay of the distribution of annual temperatures is analyzed for eight regions covering the globe from 1960 to 2022. The empirical findings reveal that extreme warming exhibits heterogeneity across regions and seasons, with a pronounced distinction between the Northern and Southern hemispheres. Notably, extreme warming predominantly occurs in both hemispheres during the period from June to September.

C1491: **The real effects of climate volatility shocks**
Presenter: **Sofia Sanchez Alegre**, University Carlos III of Madrid, Spain
Co-authors: Hernan Seoane

The real economic effects of volatility shocks related to climate change are quantified. A dynamic stochastic general-equilibrium (DSGE) closed-economy real business cycle model is developed incorporating energy consumption and the negative productivity impacts of atmospheric CO2 accumulation. The model is extended to account for the time-varying volatility of CO2 accumulation's effects on the Solow residual. The model is calibrated to match the U.S. business cycle, together with energy consumption and the evolution of the carbon stock in the atmosphere. Findings indicate that volatility shocks have significant real effects, even in the absence of changes in atmospheric CO2 concentration, as they increase the risk of climate-related economic damage. Furthermore, the effects are magnified when climate change accelerates capital depreciation.

CO010 Room BH (S) 2.05 TOPICS IN FINITE SAMPLE ECONOMETRICS Chair: Antoine Djogbenou

C0225: **Evaluating financial tail risk forecasts with the model confidence set**
Presenter: **Lukas Bauer**, University of Freiburg, Statistics and Econometrics, Germany

The purpose is to provide novel results on the finite sample properties of the model confidence set (MCS) applied to the asymmetric loss functions specific to financial tail risk forecasts, such as value-at-risk (VaR) and expected shortfall (ES). The focus is on statistical loss functions that are strictly consistent. The comprehensive simulation results show that, first, the MCS test keeps the best model more frequently than the confidence level $1-\alpha$ in most settings. Second, it eliminates a few inferior models for out-of-sample sizes of up to four years. Third, the MCS test shows little power against models that underestimate tail risk at the extreme quantile levels $p = 0.01$ and $p = 0.025$, while the power increases with the quantile level p. These findings imply that the MCS test may be suitable to narrow down a set of competing models but that it is not appropriate to test if a new model beats its competitors due to the lack of power.

C0973: **Higher-order moments of GMM estimators**
Presenter: **Paul Rilstone**, York University, Canada

Explicit representations of the second-order mean-squared error, skewness and kurtosis of standard GMM estimators are obtained. These are shown to be written as quadratic and cubic functions of the second through fourth moments of the influence functions of these estimators. Iterating on the optimal weighting matrix is shown to reduce the terms in each of these moments. However, there is no gain in the second-order moments obtained from additional iterations.

C0584: **Finite sample performance of bivariate interval regression: A simulation-based study**
Presenter: **Richard Moussa**, Ecole Nationale Superieure de Statistique et dÉconomie Appliquee, Cote d'Ivoire

The aim is to provide evidence on the finite sample performance of bivariate interval regression models on cross-sectional data. For this purpose, in addition to the parametric estimator, a copula-based estimator is introduced with the standard copulas. Estimator performance is assessed using (i) the bias, (ii) the average estimated standard error, and (iii) the empirical coverage probabilities. The Monte Carlo simulations assess the performance of the various models for sample sizes varying between 50 and 500 and the number of covariates per equation between 2 and 10 to accommodate common settings in empirical studies. Results show differences in performances by type of parameter (slope, variance, and correlation) according to the estimator, the sample size, and the number of covariates. The results provide insights for the selection of an appropriate estimation approach with respect to the empirical settings.

C0842: **Permutation tests for dyadic models**
Presenter: **Pujee Tuvaandorj**, York University, Canada

Dyadic data models are commonly employed to analyze interactions between population units, where it is essential to properly account for the cross-sectional dependence induced by the dyadic structure. Well-known examples include international trade, where each observation represents a pair of countries, and network formation, where binary variables indicate whether pairs of units share a link. Permutation tests are developed for generalized linear models with such data structure. The large sample validity of the proposed tests is established, conditions under which they remain exact in finite samples are characterized, and robustness in sparse dyads is examined. Simulations suggest better finite sample properties of the proposed permutation tests compared to the asymptotic tests. Applications to the gravity model of international trade are provided, along with a re-examination of the democratic trade hypothesis in international relations.

CO409 Room BH (SE) 2.01 ADVANCES IN CLUSTERING AND ITS APPLICATION AREAS Chair: Ioanna Papatsouma

C0324: **On decision making in cluster analysis**
Presenter: **Christian Hennig**, University of Bologna, Italy

The purpose is to discuss some of the decisions that are required when using cluster analysis in practice. These included connecting the aim of analysis to the choice of appropriate methodology, decisions at the preprocessing stage: standardization, transformation, dimension reduction, choice of distance measure, decisions regarding outliers and number of clusters, use of validation tools such as stability assessment, testing, and visualization, particularly for comparing different clustering and deciding between them.

C0654: **Bayesian clustering of complex data**
Presenter: **Stephen Coleman**, Oregon Health & Science University, United States

Clustering remains a frustratingly hard problem, and often, the tools available do not meet the requirements of a specific analysis. This can be due to a multitude of reasons, but some common examples include accounting for sources of technical variation, such as is common in data generated across multiple batches, modelling shared signals across multiple modalities in a principled fashion, and accounting for missing observations within the data. Some of the challenges of designing appropriate models for such problems are covered through a Bayesian lens, as well as the related problems of implementing and applying them. The focus is on examples from the biomedical field and systems biology.

C0827: **Model-based co-clustering: High dimension and estimation challenges**
Presenter: **Christophe Biernacki**, Inria, France
Co-authors: Christine Keribin, Julien Jacques

Model-based co-clustering can be seen as a particularly important extension of model-based clustering. It allows for a significant reduction of both the number of rows (individuals) and columns (variables) of a data set in a parsimonious manner and also allows interpretability of the resulting reduced data set since the meaning of the initial individuals and features is preserved. Moreover, it benefits from the rich statistical theory for both estimation and model selection. Many works have produced new advances on this topic in recent years, and a general update of the related literature is offered. It is the opportunity to advocate two main messages, supported by specific research material: (1) co-clustering requires further research to fix some well-identified estimation issues, and (2) co-clustering is one of the most promising approaches for clustering in the (very) high-dimensional setting, which corresponds to the global trend in modern data sets.

C1032: **Weighting games in clustering**
Presenter: **Marianthi Markatou**, University at Buffalo, United States

Conceptual and technical challenges are investigated in variable weighting and selection in cluster analysis of mixed-type data. The specific objectives of existing weighting procedures are quite different and are often selected without a discussion of competing objectives and the implications of these selection choices. The different objectives of existing weighting and variable selection techniques are described and investigated, and then the variable weighting technique MEDEA (Multivariate Eigenvalue Decomposition Error Analysis) is developed. It is shown that MEDEA weighting fills an important and previously unfilled niche in cluster analysis of mixed-type data. Its performance in numerous experiments with both synthetic and actual datasets is demonstrated.

CO250 Room BH (SE) 2.05 STATISTICAL METHODS FOR NETWORK PSYCHOMETRICS Chair: Federico Castelletti

C0804: **A DAG-probit model for Bayesian causal inference and causal structure learning from ordinal data**
Presenter: **Alessandro Mascaro**, Universitat Pompeu Fabra, Spain
Co-authors: Federico Castelletti, Augusto Fasano

Psychologists often aim to understand causal relationships between latent constructs observed only through ordinal data in the form of subjects' responses to items in tests and scales. In addition, they seek to quantify the strength of these relationships, as it may be of interest to identify optimal clinical interventions. Motivated by this, a fully Bayesian methodology is developed for causal structure learning and causal effect estimation from ordinal data. In particular, it is assumed that the causal structure underlying the latent constructs follows from a linear Gaussian structural equation model admitting a direct acyclic graph as a graphical representation. Ordinal data are then obtained via discretization of the latent variables, thus inducing a DAG-probit model. An MCMC scheme is devised to sample the joint posterior distribution of DAGs, DAG parameters, and unknown discretization thresholds, from which a Bayesian model averaging estimate of any causal effect of interest can be easily obtained. The methodology is evaluated in an extensive simulation study, and a real-data application is provided on survey data.

C0914: **Comparing ordinal Markov random fields in two independent samples with Bayesian model selection**
Presenter: **Maarten Marsman**, University of Amsterdam, Netherlands

In psychological network analysis, data are often assessed on an ordinal scale (e.g., variables assessed on a Likert scale). An ordinal Markov random field graphical model has been proposed recently to adequately analyze the network structure underlying ordinal data. Bayesian edge selection methodology is used in combination with Bayesian model averaging to assess the evidence for the inclusion or exclusion of individual edges in the estimated network structure. However, quite often, interest goes out to assess if and how the estimated network structure differs across groups (e.g., based on gender or clinical status). The ordinal Markov random field model is extended to two-group designs, and the Bayesian methodology is used to assess differences in the estimated network structure across groups.

C0985: **Large-scale Bayesian structure learning for Gaussian graphical models using marginal pseudo-likelihood**
Presenter: **Reza Mohammadi**, University of Amsterdam, Netherlands

Bayesian methods for learning Gaussian graphical models offer a comprehensive framework that addresses model uncertainty and incorporates prior knowledge. Despite their theoretical strengths, the applicability of Bayesian methods is often constrained by computational demands, especially in modern contexts involving thousands of variables. To overcome this issue, two novel Markov chain Monte Carlo (MCMC) search algorithms with a significantly lower computational cost than leading Bayesian approaches are introduced. The proposed MCMC-based search algorithms use the marginal pseudo-likelihood approach to bypass the complexities of computing intractable normalizing constants and iterative precision matrix sampling. These algorithms can deliver reliable results in mere minutes on standard computers, even for large-scale problems with one thousand variables. Furthermore, the proposed method efficiently addresses model uncertainty by exploring the full posterior graph space. The consistency of graph recovery, and extensive simulation study indicates that the proposed algorithms, particularly for large-scale sparse graphs, outperform leading Bayesian approaches in terms of computational efficiency and accuracy. The practical utility of the methods is also illustrated in medium and large-scale applications from human and mice gene expression studies. The implementation supporting the new approach is available through the R package BDgraph.

C0750: **Application of the meta-analytic Gaussian network aggregation approach to investigate flourishing across 22 countries**
Presenter: **Michela Zambelli**, Universita Cattolica del Sacro Cuore of Milan, Italy

As a multi-dimensional construct, recent work has begun to explore the interrelatedness among flourishing constituents and their contribution to the achievement and maintenance of individual well-being. Using the national representative survey data from the first wave of the global flourishing study (total N = 202,898), a meta-analytic Gaussian network aggregation (MAGNA) approach was applied using psychonetrics R package to investigate similarity and differences among the reciprocal interrelation of well-being components across 22 countries. MAGNA is based on estimating a Gaussian graphical model by aggregating over multiple datasets, from which a fixed-effect structure and a random-effects structure can be obtained. The MAGNA approach allowed for obtaining a common cross-country network of flourishing and a variance-covariance matrix of random effects that quantify the heterogeneity of edges across countries. It comments, using real data, on the potential of the MAGNA approach to address the need in social science research to aggregate multiple studies to obtain a sufficiently large sample size and increase the reproducibility and reliability of statistical inferences.

CO114 Room BH (SE) 2.10 STATISTICAL MODELS AND METHODS FOR EDUCATION I Chair: Marialuisa Restaino

C0495: **Fostering resilience at the university: Statistical models for assessing the influence of high schools and peers**
Presenter: **Isabella Sulis**, University of Cagliari, Italy
Co-authors: Silvia Columbu, Mariano Porcu, Cristian Usala

Factors influencing schools' ability to foster fairness and inclusion in higher education are examined. Translating the PISA OECD definition of resilient students in the university framework, resilient university students are identified based on their socioeconomic status and their academic performance by the end of their first year. Multiple data sources were integrated, including the MOBYSU.IT dataset, which contains the Italian National Student Archive (ANS) microdata of students enrolled in Italian universities, and the INVALSI surveys, which provide information on students' family socioeconomic conditions. A multilevel approach was adopted to assess the impact of high school and peer characteristics on the likelihood of fostering resilient behaviors at the university. Combining multilevel modelling and latent modelling approaches enabled an understanding of the role of school characteristics in creating profiles (i.e., latent classes) of schools that effectively balance quality and equity in their students' university outcomes.

C0533: **Studying gender disparities in STEM university credits distribution using quantile regression**
Presenter: **Riccardo De Santis**, University of Siena, Italy
Co-authors: Nicola Salvati, Francesco Schirripa, Antonella D Agostino

The relationship between gender and university performance of students enrolled at a 3-year STEM (Science, Technology, Engineering and Mathematics) degree is investigated. The focus is on gender differences across different quantiles. The statistical modelling of earned credits encounters challenges posed by the discrete and often irregular nature of the observed distribution. Moreover, the hierarchical structure of the data demands an estimation strategy that extends beyond the simplicity of quantile regression. A methodology is implemented based on the jittering approach for counts and on penalized fixed effects in order to deal with these two distinct extensions over standard quantile regression. Data is acquired from two administrative databases made available through an agreement with the Italian Ministry of University and Research (MIUR). In the empirical analysis, data from the cohort of 2018/2019 Italian high school graduates who enrolled in the Italian university system in the academic year 2019/2020 in a STEM field is used.

C0627: **Clustering of Italian higher education institutions based on the mobility choices of academic students**
Presenter: **Silvia Bacci**, University of Florence, Italy
Co-authors: Bruno Bertaccini, Luca Scaffidi Domianello

The present contribution investigates the dynamics of the mobility of Italian university students from the geographical area of residence (origin) to the university where the student is enrolled in a bachelor's degree program (destination). Student mobility flows are typically examined using gravity models. Gravity models generally assume a uniform relationship for each origin-destination pair. This assumption may not hold for the Italian higher education system, where, due to the decentralization process, some universities show national attractiveness while others attract mainly local people. Furthermore, the flow of students from the south of Italy to the universities located in the north is more often observed than that of the opposite. For these reasons, gravity models with both destinations are considered - and origin-specific distance parameters are used to gather detailed insights for each university and the student's province of residence. The empirical analysis based on the local deterrence effect and distinct interaction dynamics among Italian universities and provinces of origin validates our hypothesis of spatial heterogeneity. Then, a fuzzy clustering technique is proposed based on the estimated distance parameters to identify both homogeneous clusters of universities, on the one hand, and homogeneous clusters of provinces, on the other hand, sharing a similar deterrence effect.

C1035: **The mathematics performance among adolescents in the COVID-19 era**
Presenter: **Mariangela Zenga**, Universita degli Studi di Milano-Bicocca, Italy

Co-authors: Adele Marshall

The impact of the COVID-19 pandemic on the mathematical performance of adolescents is explored using data from the OECD's Program for International Student Assessment (PISA) in 2022. The pandemic has significantly affected the mental health and well-being of adolescents, disrupting their daily routines and educational activities. Using a three-level random intercept regression model, data is analyzed on math scores, considering variables at the student, school, and country levels. Results show differences in performance, influenced by well-known factors such as gender, socioeconomic status, and school environment, as well as the duration of time students stayed at home due to pandemic restrictions.

CO302 Room BH (SE) 2.12 SPATIOTEMPORAL INFECTIOUS DISEASE MODELING AND SURVEILLANCE Chair: Chawarat Rotejanaprasert

C0712: **Multi-scale spatiotemporal Covid-19 modeling: The mortality example**

Presenter: **Andrew Lawson**, Medical University of South Carolina, United States

Data resources are now available that can more fully characterize the dynamics and progression of the COVID-19 pandemic around the world. In the US, county-level weekly data is now available from the Center for Disease Control (CDC), including incident case and mortality counts (as well as cumulative counts) for 173 weeks during the pandemic period. As both state-level and county-level data are now present, multiscale models for the dynamics of the pandemic process are considered. The linkage is described between state-level variation and the dynamics of county-level mortality. A variety of models of varying complexity are evaluated. An initial example of a single state (South Carolina) is demonstrated, whereby a time series model for the state level is linked to county-level weekly space-time variation. Goodness-of-fit (WAIC, MSE, DIC) is evaluated, and model extensions are reported to include deprivation measures and predictive metrics (out-of-data MSPE).

C1108: **Breaking down homogeneous mixing assumptions in epidemic compartmental models**

Presenter: **Jorge M Mendes**, NOVA Information Management School, NOVA University Lisbon, Portugal

A spatial age-structured SIR model is proposed to refine the simplistic assumption of homogeneous mixing found in traditional compartmental models by incorporating population stratification by age groups and spatial distribution. This approach acknowledges the heterogeneity in contact patterns and mobility behaviours across different demographics and geographical locations. The model utilises a contact-mobility stochastic matrix, which integrates network analysis features such as betweenness and clustering. These network metrics significantly influence the dynamics of infection transmission by either facilitating or impeding the spread of disease. Betweenness centrality identifies key individuals or nodes that act as bridges within the network, playing a crucial role in potential outbreak propagation. Clustering reflects the degree to which nodes cluster together, affecting local transmission dynamics. By incorporating these elements, the model offers a more nuanced understanding of epidemic spread, capturing the complex interplay between social structure and disease dynamics. This spatial and age-stratified approach allows for more accurate predictions and effective intervention strategies, ultimately enhancing public health responses to infectious disease outbreaks.

C1158: **Directionally dependent spatial infectious disease models**

Presenter: **Rob Deardon**, University of Calgary, Canada

In many epidemic systems, the disease can be prone to spread in some directions more than others. This can be due to migration and behavior patterns or due to prevailing wind patterns. Individual-level models (ILMs) are commonly used for modelling spatial risk in infectious disease transmission but have not traditionally considered these directional tendencies. A class of ILMs that allow for the directional dynamics of disease transmission is introduced. In these directionally dependent ILMs, the probability of an individual being infected depends on both the direction and distance between susceptible and infectious individuals. The characteristics of these directionally dependent ILMs are discussed, and how they can be fitted in a Bayesian Markov chain Monte Carlo (MCMC) framework are shown, as well as results for both simulated and real data for crop and livestock diseases.

C1174: **A new similarity-based spatiotemporal model for Covid-19 infection prediction and forecasting**

Presenter: **Helena Baptista**, Universidade Nova de Lisboa, Portugal

Co-authors: Jorge M Mendes, Ying C MacNab

A conditionally specified Gaussian random field (CS GRF) model with a similarity-based non-spatial weight matrix to facilitate non-spatial smoothing in Bayesian disease mapping (BDM) has been proposed. The model, named similarity-based GRF, is motivated for modelling disease mapping data in situations where the underlying small area relative risks and the associated determinant factors do not vary systematically in space, and the similarity is defined by similarity with respect to the associated disease determinant factors. The method, designed to handle cases when there is no evidence of positive spatial correlation, was also used when the appropriate mix between local and global smoothing is not constant across the region. It showed again that results consistent with the published knowledge were produced and that accuracy was increased to clearly determine areas of high- or low-risk. Now, the proposed method was used when the underlying small area relative risks vary systematically in space and time. This spatiotemporal approach of the method employs a dynamic CS GRF model with a novel approach to characterize infection risk dependencies through the similarity of areal-level covariates. Furthermore, the method produces the best-balanced forecast. Once again, BDM models gain significantly in explanatory and forecasting power by including extra information, according to the specific knowledge of the epidemiologists, to fit the right suitable model to the problem at hand.

CC499 Room BH (SE) 2.09 ECONOMETRIC MODELS IN ENERGY AND FINANCE Chair: Joachim Schnurbus

C1621: **Analyzing the impact of photovoltaic self-consumption on Spanish electricity consumption patterns**

Presenter: **Eduardo Caro Huertas**, Universidad Politecnica de Madrid, Spain

Co-authors: Jesus Juan

The impact of photovoltaic (PV) self-consumption on electricity consumption patterns in Spain is investigated. With the growing adoption of PV systems, households and industries are increasingly generating their own electricity, leading to shifts in overall consumption behaviors and grid demand. The analysis utilizes public data from the Spanish Transmission System Operator (TSO), examining changes in electricity usage, peak demand reduction, and the alignment of consumption with solar radiation. Findings indicate that PV self-consumption significantly alters daily and seasonal electricity consumption patterns and reduces dependency on the grid. These insights are valuable for policymakers and energy providers in designing strategies to optimize grid operations and support the transition towards renewable energy sources.

C1525: **'One out of many': Consolidating a long-term trend forecast for investing in energy commodities**

Presenter: **Fernanda Diaz-Rodriguez**, Universidad Complutense de Madrid, Spain

Co-authors: M Dolores Robles

The aim is to add to the few studies on the annual horizon of energy price forecasting. Accurately predicting trends in energy commodity prices is vital for economic and financial stability. However, the trend itself is difficult to capture, making single forecasts unreliable. A novel approach is proposed by combining multiple trend forecasts for crude oil prices, and its effectiveness is assessed in long-term investment strategies. Combining forecasts is known to improve accuracy and reduce risk compared to relying on a single method. This approach is particularly valuable for unobservable variables like trends. Five individual trend estimation methods and seven forecast combination methods are compared. The results show that combining forecasts with a novel method called METS performs best, minimizing errors and remaining stable across market conditions. Furthermore, using estimated trends outperforms using actual prices for long-term forecasting, highlighting the limitations of short-term price predictions.

C0960: **Forecasting of intraday trading volume using Bayesian nonlinear ACV models for a VWAP trading strategy**

Presenter: **Roman Huptas**, Krakow University of Economics (Uniwersytet Ekonomiczny w Krakowie), Poland

The aim is to evaluate the accuracy of intraday volume point forecasts obtained from alternative Bayesian nonlinear autoregressive conditional volume (ACV) models in terms of a daily volume weighted average price (VWAP) trading strategy. The VWAP trading strategy is the most popular algorithmic trading strategy due to its operational simplicity. Its goal is to split large orders into smaller-sized orders and execute them during the trading day to achieve an average price that is close to the VWAP. The VWAP strategy needs to be based on accurate intraday volume point forecasts, which are crucial to accomplish its goal. More accurate predictions result in a better-executed VWAP and lower execution risk. Different specifications of ACV models are analyzed and compared: a linear, logarithmic and Box-Cox ACV model. The exponential and the generalized gamma distributions of error terms are examined. Additionally, a log-normal distribution for innovations is also explored since it has received no attention in the literature on this type of model. The ACV models are compared to benchmarks such as the naive method and the rolling means technique. Two types of VWAP replication strategies, static and dynamic, are considered, and VWAP tracking mean squared error is applied to measure the VWAP order execution risk. The empirical part includes forecasting 10-minute volume data of selected widely traded stocks from selected well-developed stock exchanges.

C1359: **Symbolic portfolios with intrinsic explainability**

Presenter: **Jonathan Chassot**, University of St.Gallen, Switzerland

Co-authors: Erik-Jan Senn

The purpose is to introduce SPIcE (Symbolic Portfolios with Intrinsic Explainability), a novel framework for portfolio optimization leveraging symbolic regression to derive interpretable expressions that guide investment decisions. Unlike traditional machine learning models, which often function as "black boxes", SPIcE provides clear and concise mathematical expressions that financial practitioners can easily understand, explain, and extend. This approach bridges the gap between the complexity of modern machine learning and the demand for transparent and interpretable models in financial decision-making. Applied to a set of real-world financial assets, SPIcE demonstrates its potential to provide portfolio managers with highly performant strategies, actionable insights, and a deeper understanding of the underlying dynamics that drive asset returns.

Saturday 14.12.2024 16:50 - 18:55 Parallel Session E – CFECMStatistics2024

CO341 Room S-2.25 DIRECTIONAL STATISTICS Chair: Anahita Nodehi

C0585: **Nonparametric collective (spectral) density estimation with applications in Bioinformatics**
Presenter: **Mehdi Maadooliat**, Marquette University, United States

A nonparametric method is reviewed for the collective estimation of multiple bivariate density functions for a collection of populations of protein backbone angles. This collective density estimation approach is widely applicable when there is a need to estimate multiple density functions from different populations with common features. An extension of this approach is then presented for the simultaneous estimation of spectral density functions (SDFs) for a collection of stationary time series that share some common features. A collective estimation approach pools information and borrows strength across the SDFs to achieve better estimation efficiency. Also, each estimated spectral density has a concise representation using the coefficients of the basis expansion, and these coefficients can be used for visualization, clustering, and classification purposes. The Whittle pseudo-maximum likelihood approach is used to fit the model, and an alternating block-wise Newton-type algorithm is developed for the computation. A web-based ShinyApp is developed for visualization, training, and learning the SDFs collectively using the proposed technique. Finally, the method is applied to cluster similar brain signals recorded by the electroencephalogram to identify synchronized brain regions according to their spectral densities.

C0628: **A versatile trivariate wrapped Cauchy copula**
Presenter: **Sophia Loizidou**, University of Luxembourg, Luxembourg
Co-authors: Christophe Ley, Shogo Kato, Kanti Mardia

A new flexible distribution for data on the three-dimensional torus is proposed, which is called a trivariate-wrapped Cauchy copula (TWCD). The trivariate copula has several attractive properties. It has a simple form of density and is unimodal. Its parameters are interpretable and allow an adjustable degree of dependence between every pair of variables, which can be easily estimated. The identifiability condition simplifies the model parameter dimension. The conditional distributions of TWCD are well studied bivariate and univariate wrapped Cauchy distributions. Furthermore, the distribution can be easily simulated, and parameter estimation is possible via maximum likelihood. Another interesting feature of this model is that it can be extended to a cylindrical copula. TWCD is illustrated on data from protein bioinformatics of conformational angles and the cylindrical copula using climate data related to a buoy in the Adriatic Sea.

C0803: **Generalized Laplace regression to model cylindrical responses with an application to physical activity in children**
Presenter: **Marco Geraci**, Sapienza University of Rome, Italy

A multivariate regression model is proposed for cylindrical responses using the (symmetric) generalized Laplace (GL) distribution. This distribution has a shape parameter that captures the heaviness of the tails and includes the Gaussian and (classical) Laplace distributions as special cases. The likelihood is obtained by first projecting the scale-mixture representation of the GL onto the circle and then by numerically integrating it with respect to the latent random variance. In an application to accelerometer data collected from participants of the UK Millennium Cohort Study, children's physical activity behaviors are studied in free-living conditions. The outcome consists of two components, one related to the intensity of the physical activity (linear and univariate or multivariate) and one related to its timing over the day (circular and univariate). In summary, the proposed model represents a useful extension of models based on the normal and the projected normal distribution.

C0966: **Addressing boundary inefficiency of local density estimators**
Presenter: **Marco Di Marzio**, University of Chieti-Pescara, Italy
Co-authors: Stefania Fensore, Chiara Passamonti

Local density estimation methods suffer severe bias near the support boundaries due to the well-known estimated density overflow problem. Because of its practical occurrence, the counterpart of such phenomenon in regression assumes strong relevance, too. The so-called reflection method appears to be an efficient way to alleviate this issue. Simple domain transformations could favor an elegant application of such a principle.

C0970: **Gaussian-related graphical models for circular variables**
Presenter: **Agnese Panzera**, University of Florence, Italy
Co-authors: Anna Gottard

Graphical models are a key class of probabilistic models for studying conditional independence among random variables. Two classes of multivariate circular distributions related, albeit in different manners, to the Gaussian distribution are considered. The main properties of these distributions are explored in terms of conditional independence, proposing related classes of graphical models. The usefulness of the proposed models is illustrated by modelling the conditional independence among dihedral angles, which are crucial for defining the three-dimensional structure of proteins.

CO041 Room S-1.01 HiTEc: CLUSTERING OF COMPLEX DATA STRUCTURES Chair: Maria Brigida Ferraro

C0922: **Mixture of generalized latent trait analyzers for jointly clustering pediatric patients and their clinical conditions**
Presenter: **Dalila Failli**, University of Florence, Italy
Co-authors: Maria Francesca Marino, Francesca Martella

Understanding how different subsets of clinical conditions manifest in pediatric patients can enhance diagnostic accuracy. The aim is to identify groups of pediatric patients possibly affected by appendicitis being similar with respect to subsets of clinical conditions. To achieve this, the finite mixture of generalized latent trait analyzers (MGLTA) is introduced, allowing to 1) handle mixed-type data; 2) group pediatric patients into distinct subsets, called components, and, within each component, identify subsets of qualitative/quantitative clinical conditions, called segments. The latter are identified via a parsimonious and flexible specification of the linear predictor. The continuous latent trait incorporated into the model allows to account for possible residual dependence between clinical conditions from the same patient. An EM algorithm is employed for the estimation of model parameters in a maximum likelihood framework, and a Gauss Hermite quadrature is considered to approximate multidimensional integrals not available in closed-form.

C1018: **Handling outliers when clustering three-way data**
Presenter: **Paul McNicholas**, McMaster University, Canada
Co-authors: Katharine Clark

A paradigm for clustering three-way data is introduced based on matrix variate mixture models. Then, an algorithm for dealing with outliers is introduced. Crucially, this algorithm does not require pre-specification of the number of outliers. The performance of the proposed approach is demonstrated using simulated and real data.

C1376: **Fuzzy clustering approaches for star-shaped sets: A comparative study**
Presenter: **Ana Belen Ramos-Guajardo**, Fundacion Universidad de Oviedo, Spain
Co-authors: Maria Brigida Ferraro, Gil Gonzalez-Rodriguez

The aim is to develop a novel method to identify groups of star-shaped sets in Rp. These sets are characterized by a central point and a radial function that accounts for directional inaccuracy around that center. To achieve this, a new fuzzy clustering algorithm based on the Mahalanobis distance is proposed, incorporating the covariance matrices associated with each cluster. The use of the Mahalanobis distance in clustering has been demonstrated to effectively identify non-spherical clusters, which are often overlooked when employing a Euclidean-type distance. A comparative analysis is conducted between this method and a previously proposed generalization of the fuzzy k-means algorithm for star-shaped sets, focusing on their performance in a real-life application involving clasts from the Cantabrian Coast.

C1585: **A comparison of parsimonious families of hidden Markov models for multivariate longitudinal data**
Presenter: **Mackenzie Neal**, McMaster University, Canada
Co-authors: Paul McNicholas

The popularization of hidden Markov models (HMMs) for analyzing multivariate longitudinal datasets wherein estimating state switching of subjects is desirable has resulted in over-parameterization issues. Thus, parsimonious HMMs are essential for the analysis of longitudinal datasets. Commonly, parsimony is introduced by imposing a series of constraints on decomposed covariance matrices. Two families of HMMs arising from the modified Cholesky decomposition and the eigenvalue decomposition on various longitudinal datasets are introduced and compared.

C1271: **Hypothesis test based document clustering**
Presenter: **Gian Mario Sangiovanni**, Sapienza University, Italy
Co-authors: Louisa Kontoghiorghes, Ana Colubi, Maria Brigida Ferraro

It is well-known in the literature that the main limitations of document clustering techniques are that they operate in a high-dimensional space, and it is difficult to interpret the different clusters once a partition has been obtained. The proposed methods for computing document clustering employ a two-stage process. Initially, it can be observed that the information contained within the document-term matrix exhibits significant sparsity, so a direct application of a clustering technique would be highly inefficient. Consequently, dimensionality reduction is applied. The proposed strategy involves employing latent Dirichlet allocation (LDA) to identify the main topics in the corpus under analysis. To determine the similarity between two documents, the p-value of a hypothesis test of the homogeneity of topic distributions between two documents is computed. This p-value is used as a similarity measure, upon which three different clustering procedures are built. The first two directly employ the new dissimilarity using a hierarchical approach and a fuzzy relational clustering approach, while the other is a test-based approach to clustering. The performance of the clustering methods is then assessed using some benchmark datasets in order to understand the advantages and disadvantages of the proposals.

CO141 Room S-1.04 BRANCHING AND RELATED PROCESSES II Chair: Ines M del Puerto

C1675: **Limit laws for tree-indexed autoregression**
Presenter: **Anand Vidyashankar**, George Mason University, United States
Co-authors: Giacomo Francisci

A branching random walk model is considered, incorporating an autoregressive structure and the point processes representing the positions. The process starts at time 0 with a single ancestor at the origin and evolves as follows: each individual produces a random number of children whose positions Z_v are displaced from their parents' position Z_{Av} by a factor of $\rho \in \mathbb{R}$, that is, $Z_v = \varepsilon_v + \rho Z_{Av}$ where Av is the parent of v. Here, $\{\sum_v \delta_{\varepsilon_v}\}$ is a collection of i.i.d. point processes and δ_x is the Dirac measure at x. The case $\rho = 1$ corresponds to the classical branching random walk. The model exhibits substantially different behavior depending on whether $|\rho|$ is smaller or larger than one. In both cases, the convergence of the rescaled Laplace transform of positions at generation n is studied, and an analog of the Kesten-Stigum theory is established for these processes. Almost sure convergence of the rescaled average positions are also established at generation n, and the related central limit theorems is derived.

C0248: **Modelling predator-prey systems with two-sex branching processes**
Presenter: **Cristina Gutierrez Perez**, University of Extremadura, Spain
Co-authors: Carmen Minuesa Abril

The purpose is to introduce a two-sex and density-dependent branching process aimed at describing predator-prey dynamics. Both populations are assumed to engage in sexual reproduction with promiscuous mating. The control over the total number of individuals from each species, which produces offspring after the mating phase in each generation, is modelled using a binomial distribution. The size of this distribution depends on the number of individuals, and the probability of success is influenced by the prey density per predator. This model allows for the characterization of typical cyclic behavior observed in real predator-prey systems. Additionally, results related to the fixation, extinction, and coexistence of both species are presented. In this last case, different growth rates are observed for the prey-to-predator ratio depending on the relation between the parameters of the model.

C1437: **Spectral methods and their application in stochastic analysis**
Presenter: **Elena Yarovaya**, Lomonosov Moscow State University, Russia

The development of spectral technique allows obtaining limit theorems on the numbers of particles in a branching random walk on points of a multidimensional lattice under the assumption of the existence of branching sources (i.e., lattice points, in which particles can multiply and die) with both positive and negative branching intensities. The results on the relationship between the structure of the evolution operator spectrum and the geometric location of branching sources on a multidimensional lattice will be presented. As a rule, in earlier studies, the underlying random walk was assumed to be symmetric. It is shown that the obtained results remain valid when the condition of self-adjointness of the operator defining the random walk to a weaker condition of similarity to the self-adjoint one. Thus, with the use of spectral techniques, problems are solved related to multipoint perturbations of operators arising in the evolution equations for the first moments of particle numbers in multitype branching random walks and proved a number of new limit theorems on the behavior of populations and subpopulations of particles in a branching random walk.

C1182: **Ancestral reproductive bias in branching processes**
Presenter: **Samuel Johnston**, Kings College London, United Kingdom

Consider a simple branching process modelling a growing population. A single particle at time zero is started with, and thereafter, each particle has a standard exponential lifetime. At the end of its lifetime, it dies and is instantaneously replaced by two particles. Allowing the process to run until a time $T > 0$ is considered, and a single particle is chosen uniformly at random from the population. The times at which the ancestors of this particle died are studied, and it is found that the reproduction law along this ancestral lineage is faster than that of the underlying population population. This is due to an "inspection paradox": cells with faster lifetimes are more likely to have one of their descendants sampled by virtue of their prolificity. This inspection paradox is explored (as well as other inspection paradoxes in the broader probability literature), and the resulting bias is linked to recent observations in genetic data.

C1524: **Asymptotic behavior of critical multitype branching processes with random migration**
Presenter: **Miguel Gonzalez Velasco**, University of Extremadura, Spain
Co-authors: Pedro Martin-Chavez, Ines M del Puerto

The aim is to introduce a multitype branching process with random migration and to study its asymptotic behavior. The focus is on what is referred to as the critical case. Sufficient conditions are provided to determine whether the process exhibits unbounded growth or not. Furthermore, by using appropriate normalizing sequences, the asymptotic distribution of the process is analyzed. Finally, a Feller-type diffusion approximation is obtained.

CO165 Room S-1.06 OVER-PARAMETRIZATION AND OVERFITTING IN MACHINE LEARNING Chair: Debarghya Ghoshdastidar

C1001: **Deep learning at smaller scale**
Presenter: **Rebekka Burkholz**, CISPA Helmholtz Center for Information Security, Germany

Deep learning continues to achieve impressive breakthroughs across disciplines but relies on increasingly large neural network models that are trained on massive data sets. Their development inflicts costs that are only affordable by a few labs and prevents global participation in the creation of related technologies. The focus is on the question of whether it really has to be like this, and some of the major challenges that limit the success of deep learning on smaller scales are discussed. Three examples of complimentary approaches are given that could help addressing the underlying issues: (i) early neural network sparsification, (ii) the integration of useful inductive bias in the design of problem specific neural network architectures, and (iii) improving training from scratch.

C0934: **Surprising learning curves: More data can lead to worse performance and worse estimators**
Presenter: **Tom Viering**, TU Delft, Netherlands

Learning curves plot performance on unseen data (such as risk) versus dataset size used for learning or estimation. It is noted that these curves are essentially different from similarly named curves that plot performance versus epoch. Learning curves can be used to estimate the amount of data needed for learning. Learning theory and regular statistical results would suggest that learning curves always improve with more data; indeed, many generalization bounds and theory indicates that the excess risk should decrease at a rate of 1/n or 1/sqrt(n), where n is the size of the dataset used for learning or estimation. In contrast, the focus is on some surprising learning curves called ill-behaved: learning curves with maxima and minima, and even periodicity, meaning that more data leads to worse performance. This happens even in expectation, even if the probabilistic model is well-specified, can happen for any sample size and can also occur in Bayesian settings. Surprising behaviors, their relation to double descent, and why they do not contradict well-understood theoretical results are discussed. It highlights less understanding about learning curves than perhaps expected and shows the need for more study of basic machine learning and statistics.

C0418: **Theoretical foundations of scaling**
Presenter: **Leena Chennuru Vankadara**, Amazon Web Services, Germany

Scaling is pivotal to the success of modern machine learning. However, this upscaling also introduces new challenges, such as increased training instability. Given the immense resources required, developing high-confidence scaling hypotheses backed by rigorous theoretical research is crucial. The purpose is to discuss how infinite width theory can be utilized to establish optimal scaling rules across various architectures and learning paradigms. It begins by discussing the scaling behaviour of multilayer perceptrons (MLPs) under sharpness aware minimization, a min-max learning formulation designed to enhance generalization. The analysis extends naturally to other architectures like transformers, ResNets, and CNNs. Additionally, the scaling behavior of structured state space models (SSMs), which have emerged as efficient alternatives to transformers, is discussed. Owing to the unique structure of their transition matrices, SSMs defy conventional scaling analyses and necessitate specialized approaches. The scaling of SSMs is discussed within the standard minimization framework, highlighting the need for and implications of specialized scaling strategies.

C0374: **When can we approximate wide contrastive models with neural tangent kernels and principal component analysis**
Presenter: **Pascal Esser**, Technical University of Munich, Germany
Co-authors: Gautham Anil, Debarghya Ghoshdastidar

Contrastive learning is a paradigm for learning representations from unlabelled data, and several recent works have claimed that such models effectively learn spectral embeddings and show relations between (wide) contrastive models and kernel principal component analysis (PCA). However, it is not known if trained contrastive models indeed correspond to kernel methods or PCA. The training dynamics of two-layer contrastive models are analyzed with non-linear activation, and it is answered when these models are close to PCA or kernel methods. It is well known in the supervised setting that neural networks are equivalent to neural tangent kernel (NTK) machines and that the NTK of infinitely wide networks remains constant during training. The first constancy results of NTK are provided for contrastive losses, and a nuanced picture is presented: NTK of wide networks remains almost constant for cosine similarity-based contrastive losses but not for losses based on dot product similarity. The training dynamics of contrastive models are further studied with orthogonality constraints on the output layer, which is implicitly assumed in works relating contrastive learning to spectral embedding. The deviation bounds suggest that representations learned by contrastive models are close to the principal components of a certain matrix computed from random features.

C0963: **Double descent and benign overfitting for error in variables regression**
Presenter: **Rishi Sonthalia**, Boston College, United States

Many prior works have used random matrix theory to understand the generalization risk for linear regression. The focus is on adding noise to both the output and the input. This change significantly affects the learning process. Using tools from random matrix theory, the generalization error is derived for low-dimensional data. Results are provided on covariate shifts and an astonishing phenomenon is noticed: double descent is obtained with under-parameterized models.

CO101 Room S-1.27 RESAMPLING METHODS IN MODERN SETTINGS Chair: Miles Lopes

C0279: **Edgeworth expansion and bootstrap for entrywise eigenvectors statistics of low-rank random matrices**
Presenter: **Fangzheng Xie**, Indiana University, United States

Understanding the distributions of spectral estimators in low-rank random matrix models, also known as signal-plus-noise matrix models, is fundamentally important in various statistical machine-learning problems, including network analysis, matrix denoising, and matrix completion. The distributions of entrywise eigenvector statistics are studied for a broad range of signal-plus-noise matrix models by establishing their Edgeworth expansion formulae. The key to the approach is a sharp higher-order entrywise eigenvector stochastic expansion. It is shown that the first-order term in the expansion is a linear function of the noise matrix, while the second-order term is a linear function of the squared noise matrix. Furthermore, under mild conditions, it is shown that Cramer's condition on the smoothness of noise distribution is not required, thanks to the self-smoothing effect of the second-order term in the eigenvector stochastic expansion. This phenomenon is unusual in low-dimensional problems. The Edgeworth expansion result is further applied to justify the higher-order accuracy of the residual bootstrap for approximating the distributions of the studentized entrywise eigenvector statistics.

C0282: **Change-point inference for high-dimensional heteroscedastic data**
Presenter: **Xiaofeng Shao**, University of Illinois at Urbana-Champaign, United States

A bootstrap-based test is proposed to detect a mean shift in a sequence of high-dimensional observations with unknown time-varying heteroscedasticity. The proposed test builds on the U-statistic-based approach, targets a dense alternative, and adopts a wild bootstrap procedure to generate critical values. The bootstrap-based test is free of tuning parameters and is capable of accommodating unconditional time-varying heteroscedasticity in high-dimensional observations, as demonstrated in the theory and simulations. Theoretically, the bootstrap consistency is justified by using the recently proposed unconditional approach. Extensions to testing for multiple change points and estimation using wild binary segmentation are also presented. Numerical simulations demonstrate the robustness of the proposed testing and estimation procedures with respect to different kinds of time-varying heteroscedasticity.

C0350: **Parametric bootstrap on networks with non-exchangeable nodes**
Presenter: **Can Minh Le**, University of California, Davis, United States

Co-authors: Zhixuan Shao

The parametric bootstrap method for networks is studied to quantify the uncertainty of statistics of interest. While existing network resampling methods primarily focus on count statistics under node-exchangeable (graphon) models, more general network statistics are considered (including local statistics) under the Chung-Lu model without node-exchangeability. The natural network parametric bootstrap is shown to first estimate the network-generating model and then draw bootstrap samples from the estimated model generally suffers from bootstrap bias. As a general recipe for addressing this problem, it is shown that a two-level bootstrap procedure provably reduces the bias. This essentially extends the classical idea of iterative bootstrap to the network setting with a growing number of parameters. Moreover, the second-level bootstrap provides a way to construct higher-accuracy confidence intervals for many network statistics.

C1031: **Bootstrap inference for least angle regression**
Presenter: **Daniel Nordman**, Iowa State University, United States
Co-authors: Karl Gregory

Least angle regression (LARS) is a well-known algorithm for variable selection in linear regression models, which provides an alternative to forward selection. However, little is known about the distributional behavior of LARS estimators, which hinders inference with LARS. The purpose is to overview some newly established distributional properties of LARS estimators, where the latter may be viewed as estimating variable importance correlations at the population level. These distributional results also provide a helpful perspective for understanding the mechanics of LARS. LARS estimators have large-sample distributional limits that may be either normal or non-normal, depending on whether estimators are capturing a true population signal or not. However, despite these complications, a bootstrap approach can be shown to validly approximate the distributional structure of LARS estimators, and, thereby, the bootstrap allows for useful inference in LARS regarding variable importance. The bootstrap method and results are illustrated in numerical studies and examples.

CO301 **Room Auditorium** NEW DEVELOPMENTS IN FINANCIAL ECONOMETRICS **Chair: Roxana Halbleib**

C0176: **Nonlinear fore(back)casting and innovation filtering for causal-noncausal VAR models**
Presenter: **Joann Jasiak**, York University, Canada

Closed-form formulas of out-of-sample predictive densities are introduced for forecasting and backcasting of mixed causal-noncausal (structural) vector autoregressive VAR models. These nonlinear and time-irreversible non-Gaussian VAR processes are shown to satisfy the Markov property in both calendar and reverse time. A post-estimation inference method for assessing the forecast interval uncertainty due to the preliminary estimation step is introduced, too. The nonlinear past-dependent innovations of a mixed causal-noncausal VAR model are defined, and their filtering and identification methods are discussed. The approach is illustrated by a simulation study, and an application to cryptocurrency prices.

C0210: **Financial market efficiency during crisis periods: A long-memory approach based on price ranges**
Presenter: **Cristina Sattarhoff**, Kiel University, Germany

The intermittency coefficient λ^2, a parameter of the multifractal random walk model of financial volatility, measures the degree of deviation from a random walk in terms of nonlinear dependence patterns of returns and other price transformations over different time horizons (minute, daily, monthly returns, etc.). The λ^2 is estimated from price range data using the scaling property of the log-variogram. After assessing the small sample properties in a Monte Carlo simulation study by comparison with well-established estimation procedures for returns, the ranges-based method is employed to track efficiency losses during 2000-2024 for nine international stock market indices using two datasets of daily and minute data, respectively. According to the results, exceptionally large λ^2 estimates, as classified with the interquartile method, mark financial crisis years throughout. Moreover, λ^2 is estimated based on 30-minute ranges using a rolling window of 3 months, and exceptionally large values are found, which persist for nearly two weeks before the major price drops during the 2020 COVID stock market crash. These results are very promising for the early detection of crisis events, and this is also the first empirical application of the ranges-based method.

C0634: **Forecast evaluation of financial tail risk: Conditional MCS**
Presenter: **Ekaterina Kazak**, University of Birmingham, United Kingdom
Co-authors: Lukas Bauer

The purpose is to address the evaluation of point forecasts for financial tail risk through the rationale for conditional model confidence set (CMCS). Financial regulations oblige financial institutions to perform stress testing, which involves evaluating risk forecasts conditional on a range of economic indicators, such as currency or monetary risk. Thus, evaluating global out-of-sample performance may obscure conditional differences that emerge under specific states, such as business cycles or volatility regimes. A prior study proposed testing for equal conditional predictive ability (ECPA) using instrumented moment conditions, while another study advanced this by approximating the conditional expected loss differential for uniform inference. These approaches, however, assume continuous loss functions over compact subsets, which may not be practical for discrete economic states or irregular loss differentials typical in tail risk evaluation. The CMCS concept diverges from traditional approaches by allowing for model selection conditional on specific states. It aims to identify sets of models for which ECPA cannot be rejected, providing a practical tool for decision-makers. Simulations and empirical applications indicate that CMCS effectively distinguishes between models with state-dependent performance, enhancing forecast reliability in stress scenarios.

C0269: **Structural modelling of dynamic networks and identifying maximum likelihood**
Presenter: **Christian Gourieroux**, University of Toronto and CREST, Canada
Co-authors: Joann Jasiak

Nonlinear dynamic models where the main parameter of interest is a nonnegative matrixcharacterizing the network (contagion) effects are examined. This network matrix is usually assumed to either have a limited number of nonzero elements (sparsity) or admit a reduced-rank Nonnegative Matrix Factorization (NMF). We follow the latter approach and develop new probabilistic NMF inference methods. We introduce a new Identifying Maximum Likelihood (IML) method for consistently estimating the identified set of admissible NMFs and derive the asymptotic distribution of this random set. We also propose a maximum likelihood estimator of the parameter matrix for a given non-negative rank and derive its asymptotic distribution and the associated efficiency bound.

C1248: **Quantile-based modeling of scale dynamics in financial returns for value-at-risk and expected shortfall forecasting**
Presenter: **Richard Luger**, Laval University, Canada
Co-authors: Xiaochun Liu

The purpose is to introduce a new approach for forecasting value-at-risk (VaR) and expected shortfall (ES) by modeling the dynamics of the conditional scale of financial returns. Focusing on downside market risks, the conditional scale is defined as the difference between key quantiles of the return distribution, which are modeled using conditional autoregressive VaR (CAViaR) specifications. VaR is obtained by estimating the left-tail quantiles of the rescaled returns, while ES is approximated by averaging these quantiles over a range of levels below the VaR threshold, providing robust and distribution-free estimates of potential extreme losses. Simulation experiments show that this approach markedly improves the accuracy of VaR and ES forecasts, particularly in scenarios involving skewness, heavy tails, and leverage effects. The method consistently outperforms several established models, including GARCH and joint VaR and ES conditional quantile models. An empirical application using daily returns of major international stock market indices further demonstrates the model's effectiveness in accurately forecasting risk over the period studied, which includes the recent COVID-19 pandemic.

CO379 Room K0.16 ADVANCES IN NETWORKS AND CAUSAL INFERENCE Chair: Michael Schweinberger

C1129: **Rates of convergence and normal approximations for estimators of local dependence random graph models**
Presenter: **Jonathan Stewart**, Florida State University, United States

Local dependence random graph models are a class of block models for network data which allow for dependence among edges under a local dependence assumption defined around the block structure of the network. Since being introduced by a prior study, research in the statistical network analysis and network science literature has demonstrated the potential and utility of this class of models. The first statistical disclaimers are provided, which provide conditions under which estimation and inference procedures can be expected to provide accurate and valid inferences. This is accomplished by deriving convergence rates of inference procedures for local dependence random graph models based on a single observation of the graph, allowing both the number of model parameters and the sizes of blocks to tend to infinity. First, the first non-asymptotic bounds are derived on the L2 error of maximum likelihood estimators, along with convergence rates, outlining conditions under which these rates are minimax optimal. Second, and more importantly, the first non-asymptotic bounds are derived on the error of the multivariate normal approximation. Together, the developed theoretical results are the first set of conditions which achieve both optimal rates of convergence and non-asymptotic bounds on the error of the multivariate normal approximation for local dependence random graph models.

C1133: **Statistical inference and eigenvector fluctuations for random graphs with infinite rank kernels**
Presenter: **Minh Tang**, North Carolina State University, United States
Co-authors: Joshua Cape

The problem of estimating the leading eigenvectors for independent edge random graphs generated from a latent position model whose link function is possibly of infinite rank is considered. Error bounds in 2 to infinity norm are derived and row-wise normal approximations for these eigenvectors. These results are applied to the two-sample testing problem in which a pair of vertices have the same latent positions. A test statistic that converges to a weighted sum of independent chi-square is proposed under the null hypothesis.

C0496: **A regression framework for studying relationships among attributes under network interference**
Presenter: **Cornelius Fritz**, Pennsylvania State University, United States
Co-authors: Michael Schweinberger, David Hunter, Subhankar Bhadra

When network data are collected, the structure of networks is often of secondary interest compared to the question of how networks affect individual or collective outcomes. The well-established class of models known as generalized linear models (GLMs), which includes linear and logistic regression, assumes that the response of a given unit depends on predictors measured on that unit but is unaffected by predictors and responses of other units. A statistical framework is introduced that captures complex and realistic dependencies among attributes and connections while retaining the virtues of GLMs. The framework helps study relationships among attributes under network interference and is applicable to binary, count-valued, and real-valued attributes. The framework is demonstrated to be amenable to scalable statistical computing based on convex optimization of pseudo-likelihoods using minorization-maximization methods. Theoretical guarantees are established based on a single observation of dependent attributes and connections, and simulation results are presented along with an application to hate speech on the social media platform X, along with the theoretical guarantees.

C1085: **Causal inference under interference with dependent outcomes due to treatment and outcome spillover**
Presenter: **Subhankar Bhadra**, Pennsylvania State University, United States
Co-authors: Michael Schweinberger, Vishesh Karwa

Causal inference is considered under interference, with dependence among outcomes arising from treatment and outcome spillover. Two statistical contributions are made. First, the direct and indirect causal effects are characterized as a function of model parameters. Second, consistency results are established along with rates of convergence for the direct and indirect causal effects. Both are the first such results when outcomes are dependent on treatment and outcome spillover. In addition, the intervention network is allowed to be random, and the probability law that governs the intervention network is inferred, providing insight into the network-generating mechanism and helping quantify the uncertainty about the network-generating mechanism. Two approaches are developed to estimate the direct and indirect causal effects: a low-rank approximation and a minorization-maximization algorithm.

C1670: **Network two-sample test for block models**
Presenter: **Oscar Hernan Madrid Padilla**, UCLA, United States

The two-sample testing problem is considered for networks, where the goal is to determine whether two sets of networks originated from the same stochastic model. Assuming no vertex correspondence and allowing for different numbers of nodes, a fundamental network testing problem that goes beyond simple adjacency matrix comparisons is addressed. The stochastic block model (SBM) is adopted for network distributions due to their interpretability and the potential to approximate more general models. The lack of meaningful node labels and vertex correspondence translates to a graph-matching challenge when developing a test for SBMs. An efficient algorithm is introduced to match estimated network parameters, allowing the proper combination and contrast of information within and across samples, leading to a powerful test. It is shown that the matching algorithm and the overall test are consistent under mild conditions on the sparsity of the networks and the sample sizes, and a chi-squared asymptotic null distribution is derived for the test. Through a mixture of theoretical insights and empirical validations, including experiments with both synthetic and real-world data, robust statistical inference for complex network data is advanced.

CO358 Room K0.18 STATISTICAL METHODS IN WEATHER FORECASTING II Chair: Sandor Baran

C0673: **Gradient-boosted conditional vine copula models for multivariate temperature forecasting**
Presenter: **David Jobst**, University of Hildesheim, Germany
Co-authors: Annette Moeller, Juergen Gross

Weather forecasting currently relies on ensemble forecasts to address uncertainties in the future atmospheric states. However, these forecasts may exhibit biases and underdispersion. Therefore, the ensemble model output statistics (EMOS) method is frequently employed to separately postprocess the ensemble forecasts for each lead time. Unfortunately, the independence assumption among the postprocessed predictive distributions of different lead times is not always fulfilled. To restore these temporal dependencies, the postprocessed univariate distributions are combined with copula-based approaches, such as the ensemble copula coupling (ECC) or the Gaussian copula approach (GCA). As an alternative approach, a conditional vine copula (CVC) model is proposed, where the coefficients of the conditional bivariate copulas are estimated via gradient-boosting. In a case study, the suggested method is compared with ECC and GCA for the multivariate postprocessing of temperature forecasts for five lead times. The results show, that CVC provides an improvement in terms of calibration and identifies temporal dependencies better than the benchmark methods. Furthermore, the approach offers valuable insights into the covariate selection for the dependence model.

C0686: **D-vine copula based probabilistic weather forecasting**
Presenter: **Annette Moeller**, Bielefeld University, Germany
Co-authors: David Jobst, Juergen Gross

Current practice in predicting future weather is the use of numerical weather prediction (NWP) models. These models are run multiple times with different initial conditions and/or model formulations to obtain an ensemble of forecasts that represents model and forecast uncertainty. The resulting ensemble forecasts typically lack calibration and need to be corrected by statistical postprocessing models. A D-vine copula-based quantile regression (DVQR) approach is proposed for ensemble postprocessing. The DVQR incorporates important predictor variables from a large set of potentially relevant ones using a sequential forward selection procedure. It is highly data-driven and allows for the adoption of more general dependence structures as the state-of-the-art ensemble model output statistic (EMOS) postprocessing model. However, the current DVQR does not explicitly allow for the accounting of additional covariate effects, e.g. temporal or spatiotemporal information. Thus, an extension of the DVQR is introduced, where the bivariate copulas are parametrized in the D-vine copula through Kendall's tau, which can then be linked to additional covariates via generalized additive models (GAMs). The performance of the GAM-DVQR is illustrated in a case study for postprocessing of 2m surface temperature forecasts. Different GAM-DVQR models are compared to the benchmark methods EMOS, its gradient-boosted extension (EMOS-GB) and basic DVQR.

C0733: **The zero degree of freedom non-central chi squared distribution for ensemble postprocessing**
Presenter: **Juergen Gross**, University of Hildesheim, Germany
Co-authors: Annette Moeller

The possible use of the zero degrees of freedom non-central chi-squared distribution as a predictive distribution for ensemble postprocessing is introduced. This distribution has a point mass at zero by definition and is thus particularly suited for weather variables naturally exhibiting large numbers of zeros, such as precipitation, solar radiation or lightning. The performance is investigated and compared to that of the censored generalized extreme value distribution and the censored and shifted gamma distribution for postprocessing 24-hour accumulated precipitation using an EMOS (ensemble model output statistics) approach with a rolling training period.

C0783: **Enhancing multivariate post-processed visibility predictions utilizing CAMS forecasts**
Presenter: **Maria Nagy-Lakatos**, University of Debrecen, Hungary
Co-authors: Sandor Baran

Modern weather forecasts increasingly use ensemble predictions for visibility, which is crucial for aviation, maritime navigation, and air quality and impacts public health. However, visibility predictions often lack the accuracy seen in other weather variables. To address this, statistical post-processing is recommended, using historical data to align predictive distributions with actual observations. Visibility is reported in discrete values by world meteorological organization standards, leading to discrete predictive distributions. Although classification algorithms can improve forecast accuracy, they might lose spatial and temporal dependencies. Whether including Copernicus Atmosphere Monitoring Service (CAMS) predictions as an additional covariate is investigated in visibility ensemble forecasts from the European Centre for Medium-Range Weather Forecasts enhances classification-based post-processing methods and preserves spatial dependence. Joint multivariate post-processing of forecasts for all 30 investigated SYNOP observation stations is performed using the two-step ensemble copula coupling and Schaake shuffle approaches, which utilize the dependence structure of the raw vector ensemble forecasts and historical observations, respectively. It is confirmed that post-processed forecasts significantly outperform raw and climatological predictions, and incorporating CAMS forecasts further improves both univariate and multivariate predictions.

C0865: **Statistical post-processing of weather forecasts using engression**
Presenter: **Sam Allen**, ETH Zurich, Switzerland
Co-authors: Xinwei Shen, Johanna Ziegel

Numerical weather prediction models produce weather forecasts that exhibit systematic biases. To yield more reliable forecasts, statistical post-processing methods are used to recalibrate the weather model output. While statistical post-processing is now well-established within operational weather forecasting suites, most post-processing methods are univariate and, therefore, do not yield coherent forecasts for multiple weather variables, time points, and spatial locations. This can be remedied by modelling the dependence structure between different variables using copulas. However, such approaches typically either lack the flexibility required to model the complex dependencies in the weather or do not allow the inclusion of additional covariates when modelling these dependencies. The proposal is to statistically post-processing weather forecasts using engression. Engression is a distributional regression technique that combines generative machine learning with pre-additive noise, resulting in a simple yet powerful post-processing framework. Engression can be applied multivariately and also exhibits desirable theoretical properties when extrapolating beyond observed data, allowing accurate forecasts to be made at spatial locations for which data is not available, for example. To demonstrate engression as a statistical post-processing method, it is compared to state-of-the-art machine learning-based methods when post-processing temperature and wind speed forecasts in Germany.

CO194 Room K0.19 STATISTICAL CHALLENGES IN INTERDISCIPLINARY BIOMEDICAL RESEARCH Chair: Sarah Weinstein

C0536: **Bayesian estimation of the survivor average causal effect for cluster-randomized crossover trials**
Presenter: **Dane Isenberg**, University of Pennsylvania, United States
Co-authors: Michael Harhay, Fan Li, Nandita Mitra

In cluster-randomized crossover (CRXO) trials with a binary intervention, groups of individuals are assigned to one of two sequences of alternating treatments. Since clusters act as their own control, the CRXO design is typically more statistically efficient than the usual parallel-arm cluster-randomized trial. CRXO trials are increasingly popular in critical care studies where the number of available clusters is generally limited. In trials among severely ill patients, researchers often want to assess the effect of treatments on secondary non-terminal outcomes, but there may be several patients who do not survive to have these measurements fully recorded. A causal inference framework is provided to address truncation by death in the setting of CRXO trials. The survivor average causal effect (SACE) estimand is targeted, a well-defined subgroup treatment effect represented via principal stratification. Structural and standard modeling assumptions are proposed to enable SACE identification and estimation within a Bayesian paradigm. The small-sample performance of the proposed Bayesian approach for the estimation of SACE is evaluated using CRXO trial data through a simulation study. The methods are applied to a two-period cross-sectional CRXO study examining the impact of proton pump inhibitors as compared to histamine-2 receptor blockers on certain non-mortality outcomes among adults requiring invasive mechanical ventilation.

C0631: **Zero-inflated latent class mixed models for characterizing longitudinal engagement patterns**
Presenter: **Nicholas Illenberger**, NYU Langone Health, United States

The digital monitoring of patient and personal health data provides unique opportunities to improve population health outcomes. Digital health applications aimed at tracking food intake and exercise over time, for example, have shown promise in reducing the risk of diabetes when used in consultation with a patient's physician. However, the reach of these applications may be limited by differential uptake across different sectors of the population. Latent class approaches may be useful in identifying different patterns of engagement with these applications, enabling researchers to tailor future developments towards participants that may be insufficiently reached. However, measuring participant engagement with these digital platforms can be complicated in the presence of zero-inflated or missing observations. A zero-inflated extension of the latent class linear mixed effects model is developed, which can be used to identify classes of engagement trajectories based not only on expected levels of engagement but also on expected probabilities of meaningful non-engagement or missingness at each time point. The proposed methodology is applied to identify longitudinal engagement patterns among users of a digital diabetes prevention program application.

C0901: **From Poisson to Bernoulli: Unlocking the finite sample properties of survival processes**
Presenter: **Benny Ren**, Stony Brook University, United States

Finite sample inference for Cox models is an important problem in many settings, such as clinical trials. Bayesian procedures allow for finite sample inference and incorporation of prior information if MCMC algorithms and posteriors are well-behaved. In addition, estimation procedures should be straightforwardly able to incorporate multilevel modeling, such as cure models and frailty models. To tackle these modeling challenges, a uniformly ergodic Gibbs sampler is proposed for multilevel Cox models and survival functions. A novel Bayesian computation procedure is outlined that succinctly addresses the difficult problem of monotonically modeling the nonparametric baseline cumulative hazard and regression coefficients. Two key strategies are developed. First, a connection between Cox models and negative binomial processes is exploited through the Poisson process to reduce Bayesian computation to iterative Gaussian sampling. Next, it is appealed to sufficient dimension reduction to address the difficult computation of nonparametric baseline cumulative hazard, allowing for the collapse of the Markov transition operator within the Gibbs sampler based on sufficient statistics. The uniformly ergodic Gibbs sampler guarantees that MCMC draws converge in total variation distance to the posterior distribution, allowing for the constrained inference of baseline hazards in finite sample settings. The approach is demonstrated using open-source data and simulations.

C0930: **Joint estimation and false discovery control of causal effects in metabolomics randomized trials**
Presenter: **Rebecca Deek**, University of Pittsburgh, United States

Randomized experiments are agreed to be the gold standard for assessing causality. Most often, their focus is on the relationship between a single outcome and exposure of interest. However, randomized trials with multiple outcomes are becoming increasingly common. The involvement of omics data, such as metabolomics, in randomized experiments, has seen expansion that is likely to grow. For example, the interest is in using randomized experiments to confirm causal relationships, previously identified in large-scale observational studies, between metabolite concentrations and modifiable risk factors or exposures such as diet and drug or antibiotic use. Accordingly, a covariate-adjusted multivariate regression model is proposed for multivariate estimation of the average treatment effects (ATEs) while also utilizing the correlation between metabolites to improve false discovery control. The utility of the procedure is demonstrated using simulation studies and data from a real randomized clinical trial of Crohn's patients. It is shown that the proposed estimator of the ATEs is more efficient than standard estimators, and the conditional calibration procedure has better false discovery control and higher power than traditional FDR correction methods.

C1559: **Digital biomarkers of Parkinson's disease using free-living accelerometry data**
Presenter: **Danni Tu**, Regeneron Pharmaceuticals, Inc., United States

Current clinical standards for tracking the progression of Parkinson's disease (PD) rely on questionnaire-based assessments that require in-clinic visits to be administered. These instruments' subjective nature and operational limitations are a major obstacle in efficient, safe, decentralized, and remote clinical trials, warranting the need for novel digital endpoints. Wearable devices are a promising approach to monitor patients continuously and objectively, particularly in people living with neurological diseases that affect motor function. The focus is on the leveraged wrist-worn accelerometry data from the National Health and Nutritional Survey (NHANES) to develop clinically-informed measures that capture and characterize physical manifestations of PD, including altered motor function (e.g., tremor), increased fatigue, and sleep disturbances. As the NHANES cohort is not specifically characterized for PD, we also proposed a novel method to identify likely PD patients and create scores approximating the disease severity. Accelerometry-based biomarkers were shown to be specific to disease status and sensitive to disease severity scores. The proposed biomarker development framework and findings in the context of PD have the potential to enhance understanding of disease progression in PD and its impact on everyday functioning, leading to novel, more sensitive, and less burdensome digital endpoints.

CO093 Room K0.20 MODEL ASSESSMENT Chair: Maria Dolores Jimenez-Gamero

C0601: **A goodness-of-fit test for geometric Brownian motion**
Presenter: **Philipp Wuebbolding**, FH Aachen University of Applied Sciences, Germany
Co-authors: Daniel Gaigall

In the functional data setting, a new goodness-of-fit test is studied for the composite null hypothesis that the data are coming from a geometric Brownian motion. Critical values are easily obtained and ensure that the test keeps the significance level in the finite sample case. In particular, the implementation of the new approach reduces computational effort. In a comprehensive simulation study, the novel test compares favorably against competitors. An obvious application is for testing financial data, whether the Black-Scholes model applies. For illustration, data examples are provided for different stock and interest rate time series.

C0688: **A goodness-of-fit test for the geometric maximum compound logistic distribution model**
Presenter: **Daniel Gaigall**, FH Aachen University of Applied Sciences, Germany
Co-authors: Ludwig Baringhaus

Based on independent copies of a bivariate random vector (M,N), with positive integer-valued component N, testing the composite hypothesis that (M,N) follows a geometric maximum compound logistic distribution model. This distributional model is of interest, for example, in hydrology, where N models the number of floods and M is the maximum flood water level during a certain time period. The geometric maximum compound logistic distribution model is characterized in the sense that a special transform of (M,N) fulfils a specific equation. A weighted integral of an expression is suggested, obtained by replacing the function part of this equation with empirical counterparts as test statistics and proposing a parametric bootstrap procedure to get critical values. A simulation study shows the performance of the new procedure. The test is applied to a hydrological data set. A new goodness-of-fit test for the logistic distribution is obtained as a special case of the novel approach.

C0899: **From independence tests to variable selection problems: Moving beyond Euclidean settings**
Presenter: **Bojana Milosevic**, University of Belgrade, Serbia
Co-authors: Jelena Radojevic

Variable screening procedures based on independence tests are considered. Special focus is placed on moving beyond the classical Euclidean setting, specifically to accommodate hyperspherical data among other types. Mixture data types are also considered. Extensive empirical studies indicate the robustness and adaptability of the proposed method to various high-dimensional structures.

C0961: **Tests for left-truncated and right-censored data**
Presenter: **Adrian Lago**, Universidade de Vigo, Spain
Co-authors: Juan-Carlos Pardo-Fernandez, Jacobo de Una-Alvarez

The comparison of distributions is not only a classical task in statistics but also a useful tool employed in many applied fields. A great variety of tests based on different quantities or functions related to a random variable have been developed for complete data. Regarding failure times, it is also common not to know precisely the time when an individual experiments with the event of interest. Such an individual is said to be censored. In addition, it may also happen that the individuals cannot be included in the study since the event of interest occurs before the observation time. This phenomenon is called left truncation. Both censoring and truncation yield biased estimators, which implies that usual inferential methods are no longer adequate. In particular, not taking into account truncation and censoring causes inconsistent tests. On top of that, the literature on tests for left-truncated and right-censored data is vaguely developed, with the well-known rank-based tests the only option to tackle such a task. The aim is to propose a test to address the comparison of populations with left-truncated and right-censored data. Its asymptotic null distribution will be studied. As an alternative, a bootstrap resampling plan will be proposed to approximate the null distribution of the test statistic. The proposed method will be studied via Monte Carlo simulations. Finally, the test will be compared to the classical log-rank test.

C0328: **The k-sample problem using Gini covariance for large k**
Presenter: **M Remedios Sillero-Denamiel**, University of Seville, Spain

Co-authors: Maria Dolores Jimenez-Gamero

Given k populations and assuming that independent samples are available from each of them, the problem of testing for the equality of the k populations is addressed. With this aim, an unbiased estimator of the Gini covariance is taken as a test statistic. In contrast to the classical setting, where k is kept fixed, and the sample size from each population increases without bounds, k is assumed to be large, and the size of each sample is small in comparison to k. The asymptotic distribution of the test statistic is stated under the null hypothesis as well as under alternatives, which allows studying the consistency of the test. Specifically, it is shown that the test statistic is asymptotically free distributed under the null hypothesis. The finite sample performance of the test based on the asymptotic null distribution is studied via simulation. The proposal is applied to a real data set.

CO148 Room K0.50 EXPERIMENTAL DESIGN: SCREENING EXPERIMENTS Chair: John Stufken

C0256: **Row-constrained supersaturated designs for high-throughput screening**
Presenter: **Byran Smucker**, Henry Ford Health Systems, United States
Co-authors: Stephen Wright, Isaac Williams, Rick Page, Andor Kiss, Surendra Bikram Silwal, Maria Weese, David Edwards

High-throughput screening is widely used across many areas of science, perhaps most prominently in drug discovery. A statistically principled approach is proposed to these screening experiments, using the machinery of supersaturated designs and the Lasso. To accommodate limitations on the number of biological entities that can be pooled in a single microplate well, a new class of row-constrained supersaturated designs is presented. A computational procedure is developed to construct these designs, the effect of the constraint is studied on design quality, and it is shown via simulation that the proposed method is statistically superior to existing methods. Results of several applications of these designs are also shown in a system measuring the inhibition of metallo-beta-lactamase.

C0834: **A Supersaturated screening design framework based on lasso support recovery**
Presenter: **Kade Young**, Eli Lilly & Company, United States

Screening experiments utilize n experimental runs to determine which of p factors drive a response. Supersaturated screening designs (SSDs) represent a screening experiment where n<p+1. In this case, the full main effects linear model cannot be uniquely estimated with ordinary least squares. Thus, it is common for some type of penalized estimation method, like the lasso, to be used to perform factor screening. A framework is developed for optimal SSDs based on maximizing the support recovery probability of the lasso, and it is shown that a compound symmetric matrix (a matrix where all off-diagonals are equal) is the ideal structure of lasso information matrices for support recovery. This ideal structure allows for the theoretical justification of why some two-level SSD criteria outperform others under certain assumptions about the signs of the effects. Additionally, a design construction algorithm is presented that balances two design criteria based on how close a given design's information matrix is to the ideal structure. The performance of these constructed designs is evaluated compared to existing SSD construction methods, and their utility and limitations are discussed.

C1188: **Two new classes of mixed-level screening designs inspired by definitive screening designs**
Presenter: **Christopher Nachtsheim**, University of Minnesota, United States
Co-authors: Bradley Jones, Ryan Lekivetz, Dibyen Majumdar

Two families of orthogonal, mixed, two- and three-level screening designs are introduced. Both classes of designs leverage the structure of definitive screening designs. The first family is a class of orthogonal, mixed-level screening designs in multiples of eight runs that provide substantial bias protection of the main effects estimates due to active two-factor interactions. The second is a family of saturated designs containing m two-level continuous factors and $m-1$ two-level categorical or continuous factors in $n = 2m$ runs, where m is greater than or equal to four. A key advantage is that these designs are available for any even $n \geq 8$. For n a multiple of four, the designs are shown to be as effective as the best two-level alternative, namely Hadamard-matrix-based designs. It is shown that the designs typically have power near one for identifying up to m active main effects when the signal-to-noise ratio is greater than 1.5.

C1088: **Implementing arrays for fault detection in R software**
Presenter: **Ulrike Groemping**, Berliner Hochschule fuer Technik, Germany

Covering arrays (CAs) have been designed for so-called system behaviour testing. They ensure that the test runs cover each level combination of any set of up to t relevant factors at least once. Signing off a system based on a fault-free pass through such a CA provides protection against faults that arise from the combination of up to t factors. In a large system, if something fails in such a CA, it may still be challenging to figure out the root cause of the failure. This gives rise to locating arrays (LAs) that are CAs extended to make sure that there are no perfect ties between different level combinations of t factors. As per existing awareness, there are no R packages for CAs or LAs. Requests for suitable orthogonal arrays for system behaviour testing have been made, while CAs or LAs would often require substantially fewer runs for that purpose. The plan is to expand R's capabilities to include CAs and LAs, including design creation and possibly also subsequent analysis, as well as sequential steps like expanding a CA with the goal of resolving ambiguities. Preliminary research into the topic is presented with this implementation goal in mind. Experts in the field of CAs and LAs will not learn anything new but are invited to contribute their advice to the project of creating an R package as useful as possible for providing practitioners with design - and possibly analysis - features around CAs and LAs.

CO257 Room K2.31 (Nash Lec. Theatre) STATISTICS AND SPORT Chair: Brigitte Gelein

C0264: **Lasso multinomial performance indicators for in-play basketball data**
Presenter: **Argyro Damoulaki**, Athens University Economic Business, Greece
Co-authors: Ioannis Ntzoufras, Konstantinos Pelechrinis

A typical approach to quantify the contribution of each player in basketball uses the plus/minus approach. Such plus/minus ratings are estimated using simple regression models and their regularized variants with the response variable, either the points scored or the point differences. To capture more precisely the effect of each player and consider specific lineups, play-by-play data are needed. The aim is to investigate the performance of regularized adjusted plus/minus (RAPM) indicators, estimated by different regularized models having as a response the points scored per possession, using 2021-2022 NBA possession data (n=322,852). Simple model-based indices are initially presented starting from the ridge regression, the standard technique in the relevant literature. The lasso approach is proceeded with, which has specific advantages and better performance than ridge when compared with selected objective validation criteria. Then, regularized logistic regression models are implemented to obtain more accurate indicators since the response is a discrete variable, taking values mainly from zero to three. The final proposal is an improved RAPM measure, which is based on the expected points of a multinomial logistic regression model where each player's contribution is weighted by his participation in the team's possessions. The proposed indicator, called weighted expected points (wEPTS), outperforms all other RAPM measures we investigate.

C0280: **Drifting Markov models in learning of climbing**
Presenter: **Nicolas Vergne**, University of Rouen Normandy, France
Co-authors: Emmanouil-Nektarios Kalligeris, Vlad Stefan Barbu, Guillaume Hacques, Ludovic Seifert

The climbing dynamics of learning are investigated on a long-time scale by using drifting Markov models. Climbing constitutes a complex decision-making task that requires effective visual-motor coordination and exploration of the environment. Drifting Markov models is a class of constrained heterogeneous Markov processes that allow the modeling of data that exhibit heterogeneity. By applying the later models to real-world visual motor skill data, the aim is to uncover the persistent dynamics of learning in climbing. To that end, a real case study is conducted based on an experiment, with results that (i) Help in the understanding of skill acquisition in physically demanding environments; and (ii) Provide insights into the role of exploration and visual-motor coordination in learning.

C0535: **Investigating swimming technical skills by a double partition clustering of multivariate functional data from IMU sensor**
Presenter: **Antoine Bouvet**, University Rennes 2, France
Co-authors: Matthieu Marbac, Salima El Kolei

Investigating the technical skills of swimmers is a challenge for performance improvement that can be achieved by analyzing multivariate functional data recorded by inertial measurement units (IMU). To investigate the technical levels of front-crawl swimmers, a new model-based approach is introduced to obtain two complementary partitions reflecting each swimmer's swimming pattern and its ability to reproduce it. Contrary to the usual approaches for functional data clustering, the proposed approach also considers the information of the error terms resulting from the functional basis decomposition. Indeed, after decomposing into a functional basis with a finite number of elements, both the original signal (measuring the swimming pattern) and the signal of squared error terms (measuring the ability to reproduce the swimming pattern), the method fits the joint distribution of the coefficients related to both decompositions by considering dependency between both partitions. Modeling this dependency is mandatory since the difficulty of reproducing a swimming pattern depends on its shape. Moreover, a sparse decomposition of the distribution within components that permits a selection of the relevant dimensions during clustering is proposed. The partitions obtained on the IMU data aggregate the kinematical stroke variability linked to swimming technical skills and allow relevant biomechanical strategies for front-crawl sprint performance to be identified.

C0657: **Optimizing game data: Efficient tagging and analytics for basketball**
Presenter: **Mirko Carlesso**, University of Brescia, Italy

In the game of basketball, having effective data tagging tools and generating actionable insights from this data by means of statistical analyses is crucial for strategic decision-making and performance improvement. The aim is two-fold: (1) first, the focus is on an innovative data tagging application designed to optimize the capture of game data, balancing the trade-off between information reward and tagging time. The application aims to minimize the time required for data tagging while maximizing the analytical value derived from the tagged data, offering relevant benefits for basketball analytics; (2) second, how the use of statistical tools allows extracting actionable insights for coaches from the collected data is highlighted. The workflow encompasses an introduction to basketball analytics, a brief presentation of the data tagging application, and a comprehensive statistical analysis of the tagged data, demonstrating how coaches can leverage this information to enhance team performance and strategy.

C1371: **Statistical models for handball**
Presenter: **Dimitris Karlis**, RC Athens University of Economics and Business, Greece
Co-authors: Marius Oetting, Rouven Michels

Handball has received growing interest during the last years, including academic research on many different aspects of the sport. On the other hand, modelling the outcome of the game has attracted less interest mainly because of the additional challenges that occur. For example, it has been observed that the number of goals scored by each team is under-dispersed relative to a Poisson distribution and hence, new models are needed for this purpose. The aim is to present different models applicable to handball. To start with, models based on the Skellam distribution and its bivariate extensions can be the basis for avoiding the under-dispersion issue. Then, new models are developed for counting data with under-dispersion. Data from the German Bundesliga are used to show the potential of the new models.

CO375 Room K2.40 RECENT APPROACHES TO ENVIRONMENTAL AND SPATIO-TEMPORAL STATISTICS Chair: William Kleiber

C0647: **Neural methods for likelihood-free inference in spatial and spatiotemporal models**
Presenter: **Matthew Sainsbury-Dale**, King Abdullah University of Science and Technology, Saudi Arabia
Co-authors: Andrew Zammit Mangion, Jordan Richards, Raphael Huser

Methods for making inferences on parameters in statistical models are often based on the likelihood function. However, for many models, the likelihood function is unavailable or computationally intractable. The use of neural networks is discussed to facilitate fast, likelihood-free inference. These methods are "amortized" in the sense that once the neural network is trained with simulated data, inference from observed data is (typically) orders of magnitude faster than conventional approaches. The methodology is illustrated using spatial Gaussian and max-stable processes, and an application to a data set of global sea-surface temperature is showcased. There, the parameters of a Gaussian process model are estimated in 2161 spatial regions, each containing thousands of irregularly-spaced data points, in just a few minutes with a single graphics processing unit.

C0715: **Uncertainty quantification of spatiotemporal tensor completion**
Presenter: **Hu Sun**, University of Michigan, Ann Arbor, United States
Co-authors: Yang Chen

Tensor data, or multi-dimensional array, is a data format popular in multiple fields, such as social network analysis, recommender systems, and brain imaging. It is not uncommon to observe tensor data containing missing values, and tensor completion aims to estimate the missing values given the partially observed tensor. Sufficient efforts have been spared on devising scalable tensor completion algorithms but few on quantifying the uncertainty of the estimator. The uncertainty quantification (UQ) of tensor completion is nested under a split conformal prediction framework, and the connection of the UQ problem to a problem of estimating the missing propensity of each tensor entry is established. A novel tensor Ising model, parametrized by a low-rank tensor parameter, is introduced to model the locally-dependent data missingness, which is common for spatiotemporal tensor data. Estimating the tensor parameter is proposed by maximum pseudo-likelihood estimation (MPLE) with a Riemannian gradient descent algorithm. Extensive simulation studies have been conducted to justify the validity of the resulting conformal interval. The method is applied to the regional total electron content (TEC) reconstruction problem in geophysics.

C0789: **A geometric approach to extreme value theory with application to flood risk modelling**
Presenter: **Lambert De Monte**, University of Edinburgh, United Kingdom
Co-authors: Ioannis Papastathopoulos, Ryan Campbell, Haavard Rue

Extreme value theory (EVT) is a branch of probability and statistics concerned with the estimation of the probability (and the characterization of the behavior) of rare events of possibly unobserved magnitudes. A new methodology is discussed from the recently emerging geometric EVT framework and its application to the modelling of ocean levels exceeding a dyke in Newlyn, England. The approach exploits new theory results for the limiting distribution of the radial part of a radial-angular decomposition of a multivariate random vector of interest and for the distribution of directions in the multivariate space along which radial extremes occur. The framework allows for a natural extension of the concept of univariate return levels to multivariate settings, and the proposed statistical inference methodology offers great flexibility in modelling various (extremal) dependence structures.

C0879: **Anisotropic Gaussian random fields with identifiable parameters and penalized-complexity priors**
Presenter: **Liam Llamazares**, University of Edinburgh, United Kingdom
Co-authors: Finn Lindgren, Jonas Latz

Gaussian random fields (GFs) are key in spatial modeling and can be represented efficiently as solutions to stochastic partial differential equations

(SPDEs). These SPDEs depend on specific parameters which can be estimated using Bayesian inference. However, likelihood often provides limited insights under in-fill asymptotics, necessitating the use of priors. A parameterization of a non-stationary GF is introduced via its correlation length and diffusion matrix and constructs penalized complexity priors. Both the stationary case, where the parameters are constant in space, and the non-stationary case, where the parameters vary across the domain are addressed.

C1676: **Bayesian source apportionment of PM2.5 species data collected over space and time**
Presenter: **Veronica Berrocal**, University of California, Irvine, United States

The health burden associated with particulate matter exposure is now well-documented and recognized. Recently, various research efforts have been undertaken to better understand the health impacts of individual components of fine particulate matter, or PM2.5. While evidence in this regard is still forming, it appears clear that some of the species of PM2.5 are particularly noxious to human health. Thus, the reduction of PM2.5 species concentration is fundamental to protecting public health. To ensure that mitigation efforts are well-directed and more effective, it is important to understand what are the major sources of PM2.5 species. A Bayesian hierarchical model is proposed to perform source apportionment of six PM2.5 components observed in California in the year 2021. Adopting a functional framework, the observed log concentration of the 6 PM2.5 components is represented using a Bayesian latent factor model with site-specific latent source profiles modeled to be spatially dependent and linked to a global profile. The proposed model is applied to simulated data, successfully retrieving the pollution sources and the contribution of the sources to each pollutant. The model is also applied to the observed concentration of aluminum, sulfur, organic carbon, elemental carbon, nitrate and sulfate in California during the year 2021, identifying 4 major sources.

CO408 Room K2.41 MODERN STATISTICAL INFERENTIAL METHODS FOR COMPLEX DATA ANALYSIS Chair: Zhe Fei

C0419: **Time-aware knowledge representations of dynamic objects with multidimensional persistence**
Presenter: **Yuzhou Chen**, University of California, Riverside, United States

Learning time-evolving objects such as multivariate time series and dynamic networks requires the development of novel knowledge representation mechanisms and neural network architectures, which allow for capturing implicit time-dependent information contained in the data. Such information is typically not directly observed but plays a key role in the learning task performance. A new approach to a time-aware knowledge representation mechanism is proposed that notably focuses on implicit time-dependent topological information along multiple geometric dimensions. In particular, a new approach is proposed, named temporal multi-persistence (TMP), which produces multidimensional topological fingerprints of the data by using the existing single parameter topological summaries. The main idea behind TMP is to merge the two newest directions in topological representation learning, that is, multi-persistence, which simultaneously describes data shape evolution along multiple key parameters, and zigzag persistence to enable extracting the most salient data shape information over time. Theoretical guarantees of TMP vectorizations are derived, and its utility is shown in its application to forecasting on Ethereum blockchain datasets, demonstrating competitive performance, especially in scenarios of limited data records.

C0508: **High-dimensional vector autoregression with common response and predictor factors**
Presenter: **Xiaoyu Zhang**, Tongji University, China

The reduced-rank vector autoregressive (VAR) model can be interpreted as a supervised factor model, where two-factor modellings are simultaneously applied to response and predictor spaces. A new model called vector autoregression is introduced, with common response and predictor factors, to explore the common structure between the response and predictors in the VAR framework further. The new model can provide better physical interpretations and improve estimation efficiency. In conjunction with the tensor operation, the model can easily be extended to any finite-order VAR model. A regularization-based method is considered for the high-dimensional estimation with the gradient descent algorithm, and its computational and statistical convergence guarantees are established. For data with pervasive cross-sectional dependence, a transformation for responses is developed to alleviate the diverging eigenvalue effect. Moreover, additional sparsity structure is considered in factor loading for the case of ultra-high dimensions. Simulation experiments confirm the theoretical findings, and a macroeconomic application showcases the appealing properties of the proposed model in structural analysis and forecasting.

C0547: **ImgKnock: Knockoff inference with fundus images for glaucoma diagnosis**
Presenter: **Zhe Fei**, UC Riverside, United States

ImgKnock is an innovative pipeline that leverages knockoff inference and deep learning for the analysis of optic disc images in glaucoma. Knockoff inference, known for its ability to control false discovery rates in high-dimensional data, is adapted here to handle high-resolution medical images. ImgKnock uniquely generates latent knockoff features of fundus photographs, enabling the prediction of glaucoma-related outcomes and the identification and testing of important image features for accurate diagnosis. The method extends knockoff generation to non-tabular data, maintaining key features like the swapping property and ensuring robust feature selection. The pipeline consists of the following: i) Latent feature learning from the image data; ii) Knockoff generation on the latent features; iii) Knockoff feature selection with FDR control; i) Interpreting important features for glaucoma detection. This approach facilitates precise inference of feature importance, which is crucial for understanding the disease. ImgKnock has been successfully applied to various image datasets, including MNIST and CIFAR-10, demonstrating its versatility. ImgKnock's potential to contribute to improved diagnostic methods and a deeper understanding of the disease is highlighted.

C0732: **High-dimensional nonconvex penalized regression and post-selection least squares: A local asymptotic perspective**
Presenter: **Xiaoya Xu**, Shenzhen Polytechnic University, China

In the realm of high-dimensional linear regression, nonconvex penalized estimators have enjoyed increasing popularity due to their much-acclaimed oracle property, which holds under assumptions weaker than those typically required for convex penalized estimators to enjoy the same property. However, the validity of such oracle property of nonconvex penalization and the accompanying inference tools is questionable in the presence of many weak signals and/or a few moderate signals, which may incur substantial biases. To address this issue, a more holistic assessment of the selection and convergence properties of nonconvex penalized estimators is first provided from a local asymptotic perspective under a framework that accommodates many weak signals and heavy tail conditions on covariates and random errors. It is then shown that post-selection least squares estimation has the beneficial effect of removing the bias incurred by nonconvex penalization of moderate signals. Post-selection least squares estimators acquire convergence properties that are more desirable than nonconvex penalized estimators and, in the case of multiple solutions to the nonconvex optimization program, are rate-wise more robust against the choice of selected sets. Empirical results obtained from large-scale simulation studies corroborate our theoretical findings. In particular, the post-selection least squares method improves nonconvex penalized estimation, especially under heavy-tailed settings.

C0449: **Covariate adjustment in randomized block experiments and rerandomized experiments**
Presenter: **Yuehan Yang**, Central University of Finance and Economics, China

Blocking, a special case of rerandomization, is routinely implemented in the design stage of randomized experiments to balance the baseline covariates. A series of regression adjustment methods are proposed to efficiently estimate the average treatment effect in randomized block experiments with low- and high-dimensional covariates. The asymptotic properties of the proposed estimators are derived, and the conditions under which this estimator is more efficient than the unadjusted one are outlined. A conservative variance estimator is provided to facilitate valid inferences. The framework allows one treated or control unit in some blocks and heterogeneous propensity scores across blocks, thus including paired experiments and finely stratified experiments as special cases. Rerandomized experiments and a combination of blocking and rerandomization are further accommodated. Moreover, the analysis allows both the number of blocks and block sizes to tend to infinity, as well as heterogeneous treatment effects across blocks, without assuming a true outcome data-generating model. Simulation studies and two real-data analyses demonstrate the advantages of the proposed method.

CO089 Room S0.11 ADVANCES IN DATA INTEGRATION FOR LARGE-SCALE OBSERVATIONAL STUDIES Chair: Andrew Chen

C0294: **Modern neuroimaging data harmonization methods for multi-center studies**

Presenter: **Russell Shinohara**, University of Pennsylvania, United States

While magnetic resonance imaging (MRI) studies are critical for the diagnosis, monitoring, and study of a wide variety of diseases, their use in quantitative analysis can be complex. An increasingly recognized issue involves the differences between MRI scanners that are used in large multi-center studies. To address this, the current state of the art is to "regress out" or "adjust for" scanner differences. The field has found these methods to be insufficient, and recent developments are explored to address unique imaging data structures and analytic goals for which specialized harmonization methods are crucial. Available tools are further compared and contrasted for several common neuroimaging data settings.

C0448: **Promises of covariance harmonization in multi-site neuroimaging studies**

Presenter: **Jun Young Park**, University of Toronto, Canada

It is well-known that batch effects severely reduce data quality in modern neuroimaging or genomic studies. In response, many statistical methods have been developed in the past decade to homogenize batch effects for effective downstream statistical analyses, exemplified by the ComBat method that models heterogeneity in means and variances. It is shown that modeling heterogeneity in covariances (in addition to means and variances) substantially improves the quality of batch effect correction in neuroimaging studies. At the same time, there are multiple approaches to model covariance heterogeneity, which motivates a solid understanding of data characteristics for a more successful harmonization. RELIEF and SAN are showcased, which use a low-rank factor model or a spatial Gaussian process to model covariance batch effects in various neuroimaging data types, such as diffusion tensor imaging (DTI) or cortical thickness. The empirical performance of these methods are demonstrated using the SPINS study and discuss a few possible extensions of these methods.

C0700: **Integration of PET imaging studies in Alzheimer's disease**

Presenter: **Dana Tudorascu**, University of Pittsburgh, United States

Multisite imaging studies increase statistical power and enable the generalization of research outcomes. However, the variety of imaging acquisition scanners, different tau PET tracer properties, and inter-scanner variability hinder the direct comparability of multi-scanner PET data. Furthermore, the PET imaging field is lagging behind in terms of harmonization methods due to these challenges. This is further complicated due to the off-target binding of tau tracers, which can lead to bias in the PET outcomes and, in consequence, artificially increase or decrease effect sizes in clinical studies of Alzheimer's disease. Data integration methods and their downstream effects are investigated on effect size and sample size variability differences in two large multisite studies of Alzheimer's disease, and those are compared with PET brain phantom studies from different scanners.

C0763: **Bayesian variable selection for interval-censored outcomes in Genome-wide association studies**

Presenter: **Jaihee Choi**, Marquette University, United States

Co-authors: Ryan Sun

With the growing popularity of genetic and health databases such as the UK Biobank, there is increased access to Genome-wide association studies (GWAS) with interval-censored time-to-event outcomes. Gene set-based association tests have proven to be successful in identifying genes or risk loci associated with outcomes of interest while maintaining sufficient statistical power. However, fine-mapping the specific SNP or SNPs within these gene sets associated with the disease can lead to a better understanding of the genetic etiology of the disease. Though using interval-censored time-to-event outcomes can provide more information about the genetic pathology behind disease more than the binary or right-censored representation of the data, there currently are not many methods that work with interval-censored outcomes. A Bayesian framework is investigated for fine-mapping individual genetic variants associated with interval-censored data. This framework is applied to colorectal cancer data from the UK Biobank.

C0991: **Methods for robust multi-study genomic data integration: Applications in infectious diseases research**

Presenter: **Evan Johnson**, Rutgers University, United States

Despite widespread efforts to study the etiology of many prevalent infectious diseases, there exists broad heterogeneity in disease etiology driven by the diverse elements of host immune function and the myriad of potential disease-causing organisms. Several limiting constraints complicate the ability to use these existing data. One important challenge is the lack of molecular datasets with a large enough sample size to explore disease relationships and adequately train and validate new biomarkers. Another challenge is the lack of sufficient computational tools and platforms for data analysis and biomarker generation across multiple studies and cohorts. Resources for infectious disease biomarker generation and validation and data integration are discussed. This includes curated data platforms that provide existing gene expression data from public repositories with user interfaces to enable the interactive analysis of specific infectious diseases. Multi-study analytics and visualization are discussed, focused on ensembles of new and existing ID-related immune signatures. Novel machine learning methods are proposed for multi-study learning and unify the proposed resources to drive next-generation multi-study biomarkers for multiple infectious diseases. These data science tools and resources will accelerate research, leading to better mechanistic understanding and new biomarkers for infectious disease progression and heterogeneity.

CO115 Room S0.12 COPULA AND EXTREME DEPENDENCE Chair: Giorgia Rivieccio

C0596: **Two-sample testing for tail copulas with an application to equity indices**

Presenter: **Umut Can**, University of Amsterdam, Netherlands

Co-authors: Roger Laeven, John Einmahl

A novel, general two-sample hypothesis testing procedure is established for testing the equality of tail copulas associated with bivariate data. More precisely, using a martingale transformation of a natural two-sample tail copula process, a test process is constructed, which is shown to converge to a standard Wiener process under the null hypothesis. Hence, a myriad of asymptotically distribution-free two-sample tests can be obtained from this test process. The good finite-sample behavior of the procedure is demonstrated through Monte Carlo simulations. Using the new testing procedure, no evidence of a difference in the respective tail copulas is found for pairs of negative daily log returns of equity indices during and after the global financial crisis.

C0621: **Impact of exogenous factors on tail risk measures in Australian electricity markets**

Presenter: **Vincenzo Candila**, University of Salerno, Italy

Co-authors: Antonio Naimoli

Applying risk management strategies to financial markets is a common practice nowadays. Nonetheless, forecasting value-at-risk (VaR) and expected shortfall (ES) in the electricity markets is a relatively new area of research. The electricity markets have unique features like inelastic demand, price jumps, high volatility, strong intraday and daily seasonality, severe skewness and kurtosis, and negative prices. Moreover, the increasing electricity market liberalization has created incentives to realize active market risk management. In this context, the literature has not investigated the potential impact of low-frequency covariates, which could influence the daily VaR and ES measures. From this perspective, many price drivers, such as climate and economic policy uncertainty, could be considered. A large set of parametric, semi-parametric, and non-parametric models are used to forecast the VaR and ES of five Australian spot electricity markets, considering low- and high-frequency exogenous variables.

C0699: **Adaptive thresholding and tail index estimation under normal and extreme regimes**

Presenter: **Omid Ardakani**, Georgia Southern University, United States

Heavy-tailed and high-dimensional data present significant challenges for modeling and risk assessment. The application of Bayesian nonparametric methods and generalized Pareto distributions are extended in the context of extreme value theory and high-dimensional settings. First, a Dirichlet process mixture model is developed to estimate thresholds and tail indices. Subsequently, the challenges in high-dimensional models are addressed, focusing on extremal dependence by incorporating a tail adaptation mechanism and copula modeling. The empirical study examines the behavior of macroeconomic variables under normal and extreme regimes using these techniques to offer insight into the tail risks associated with macroeconomic outcomes.

C0906: **US banking returns and copper global price: A DCC-GARCH-MIDAS approach**

Presenter: **Michele Mario Ippolito**, University of Naples Parthenope, Italy

Co-authors: Giovanni De Luca, Giorgia Rivieccio

The purpose is to link the long-run correlation component of daily US banking returns to monthly copper global prices using the dynamic conditional correlation-MIDAS (DCC-MIDAS) framework. The DCC-MIDAS connects the DCC specification to the GARCH-MIDAS framework, where a short-run term is a GARCH component that moves around a long-run term driven by copper global prices computed over a monthly period. Findings highlight a strong negative influence of copper global price on US banking returns long-run volatility improving forecasts. The contribution to the literature is twofold. Firstly, taking into account the monthly copper price, which is heavily influenced by several structural breakdowns, the volatility of US banking returns is effectively depicted. Secondly, a measure of the dynamic correlations is provided, also in the tails, of the short-run US banking returns around the time-varying long-run correlations.

C0932: **Tail maximal dependence in bivariate models: Estimation and applications**

Presenter: **Chen Yang**, Icahn School of Medicine at Mount Sinai, United States

Assessing dependence within extreme co-movements of financial instruments has been of much interest in the domain of risk management. Typically, indices of tail dependence are used to quantify the strength of such dependence, although many of them underestimate the strength. To address this issue, the tail order of maximal dependence (TOMD) is proposed to improve the diagonal-based tail dependence indices. However, TOMD has so far lacked empirical estimators and statistical inference results, thus hindering its practical use. For this reason, a statistical procedure is developed to estimate the indices. The proposed procedure is evaluated through simulation studies, and it is further demonstrated with real-world financial data.

C0947: **A mixture copula model for assessing net bubble values risk**

Presenter: **Andrea Montanino**, University of Naples Parthenope, Italy

Co-authors: Giovanni De Luca

The Phillips, Shi and Yu (PSY) test is applied to investigate the presence of bubbles in the prices of cryptocurrencies and technology stocks. This methodology is based on a rolling window backward expansion procedure and is designed to allow the date stamping of the exuberant periods. The net bubble values are calculated as the difference between the Backward Supremum ADF test statistics and its critical value, estimated using a bootstrap procedure with 1000 iterations. Furthermore, the common trading days between financial assets are considered, allowing for a detailed analysis of trading patterns and synchronization among different markets. The linear correlation and tail dependence measures (through the copula function) are applied to analyze the behavior in the tails. In particular, the mixture copula is a flexible method to capture the complex dependencies in the tail regions of the distributions, which are critical to understanding the extreme co-movements and the risk of contagion. This methodology offers valuable insights for investors and policymakers, enabling them to better understand the complex dynamics of financial markets and to make informed decisions to mitigate the risks associated with explosive financial markets.

CO325 Room S0.13 BAYESIAN AND COMPUTATIONAL METHODS FOR BETTER HEALTHCARE DECISIONS Chair: Fan Bu

C0425: **Principal stratum strategy for safety evaluation**

Presenter: **Veronica Ballerini**, University of Florence, Italy

Co-authors: Alessandra Mattei, Fabrizia Mealli

Safety evaluation of new therapies is an essential aspect of clinical trials, primarily quantifying the incidence of adverse events (AEs) and comparing it to a standard treatment. Despite its importance, safety analysis of adverse events is often rather simplistic: AE probabilities are estimated without explicitly defining the target causal comparison and neglecting assumptions on the censoring mechanisms, leading to differential follow-up times. A first proposal is made addressing the evaluation of drug safety in the estimand strategy framework under the principal stratification approach. Principal estimands of interest are defined; among them, the adherent incidence rate is introduced, namely the time-varying incidence rate of adverse events in the (time-varying) principal stratum of potential adherent patients. The principal effects are estimated under the assumption of principal ignorability leveraging a fully Bayesian model.

C0520: **Bayesian hierarchical methods for modeling individual level variances for predicting health outcomes**

Presenter: **Irena Chen**, Max Planck Institute for Demographic Research, Germany

Longitudinal biomarker data and cross-sectional outcomes are routinely collected in modern epidemiology studies, often with the goal of informing tailored early intervention decisions. Most existing methods focus on constructing predictors from mean marker trajectories. However, subject-level biomarker variability may also provide critical information about disease risks and health outcomes. Current literature does not provide statistical models to investigate such relationships with valid uncertainty quantification. A family of Bayesian hierarchical models are developed that estimates subject-level means, variances, and co-variances of longitudinal biomarkers and uses these as predictors within a joint modeling setting. These methods are designed to handle varying levels of data complexity, such as multiple marker trajectories, repeatedly measured and cross-sectional outcomes, and individual time-varying (co-)variances. Advances in personalized healthcare are supported by modeling the complex interplay between biomarker means and variances, as well as corresponding health outcomes.

C0738: **Spatial nested mixture models for MALDI-MSI image segmentation**

Presenter: **Francesco Denti**, University of Padua, Italy

Mass spectrometry imaging is emerging as a valuable tool for measuring in-situ cancer biomarkers, as it allows the detection of the critical biological traits that would be overlooked with a simple visual morphological assessment of a sample. This technique measures the abundance of several specific molecules over multiple locations of a biological sample. The analysis of these complex data structures calls for developing tailored statistical methods. Over the last few years, the Bayesian community has dedicated increased attention to mixture priors inducing nested random partitions. Employing models for nested, separate exchangeable data is proposed to estimate a biclustering solution, i.e., cluster locations characterized by similar abundance profiles. This way, molecules can be simultaneously detected with similar expressions within clusters of pixels. Moreover, a hidden Markov random field prior is employed to perform appropriate image segmentations. To address the large dimensionality of these datasets and the need for timely results, an efficient coordinate ascent variational inference algorithm is applied that dramatically scales the model's applicability. The estimated biclustering structure is showcased, allowing the detection of meaningful image segmentation and patterns of activated molecules.

C0765: **Hamiltonian Monte Carlo for Bayesian nonparametric clustering via soft multinomial approximations**
Presenter: **Shounak Chattopadhyay**, University of California, Los Angeles, United States
Co-authors: Marc Suchard

Bayesian nonparametric (BNP) clustering approaches provide an elegant yet flexible framework to carry out model-based clustering. Posterior computation in such models typically involves Markov chain Monte Carlo (MCMC) algorithms such as Gibbs sampling, iteratively sampling the cluster membership variables and the model hyperparameters. Although straightforward to implement, Gibbs sampling can have poor computational scalability, particularly when the discrete cluster allocations exhibit mutual dependence. As an alternative, model-based soft-multinomial clustering is developed via a continuous approximation to the discrete multinomial variable signifying cluster membership. The use of soft-multinomials allows the implementation of general-purpose MCMC algorithms such as Hamiltonian Monte Carlo (HMC) to jointly sample the cluster memberships and the model hyperparameters. Theoretical results are provided detailing this approximation and demonstrate substantial computational improvements using soft-multinomial clustering with HMC for posterior sampling over discrete BNP clustering approaches in a variety of simulation examples and real-world data applications.

C1427: **Statistical learning in high-dimensional methylation data in cancer using trans-dimensional hidden Markov models**
Presenter: **Farhad Shokoohi**, University of Nevada Las Vegas, United States

The analysis of high-dimensional methylation data is increasingly critical in biology and health sciences due to its significant role in cancer development and progression. Various statistical methods and analytical tools have been developed to investigate DNA methylation, particularly in identifying differentially methylated cytosines or regions between groups, such as cancerous versus healthy tissues. However, analyzing high-dimensional methylation data presents substantial challenges, including heavy missing data, low read depths, functional autocorrelation patterns, the presence of multiple covariates, and the need to address multiple comparisons. These challenges are explored, and an overview of current methodologies and tools is provided, including two of the recently published approaches. Furthermore, a novel method that leverages trans-dimensional Markov chain Monte Carlo techniques with hidden Markov models and binomial emissions tailored for bisulfite sequencing data is introduced. The effectiveness of these methods is illustrated through both simulations and real data analyses on acute Leukemia and colorectal cancer. Insights are offered into the latest advancements in high-dimensional methylation analysis, and how these approaches can enhance the understanding of epigenetic changes in cancer are discussed, potentially leading to new therapeutic strategies.

CO391 **Room Safra Lec. Theatre** **ADVANCES IN FUNCTIONAL AND SPATIAL DATA ANALYSIS** **Chair: Anna Calissano**

C1563: **Variograms for Kriging and clustering of spatial functional data with phase variation**
Presenter: **Sebastian Kurtek**, The Ohio State University, United States
Co-authors: Karthik Bharath, Xiaohan Guo

Spatial, amplitude and phase variations in spatial functional data are confounded. Conclusions from the popular functional trace-variogram, which quantifies spatial variation, can be misleading when analyzing misaligned functional data with phase variation. To remedy this, a framework is described that extends amplitude-phase separation methods in functional data to the spatial setting, with a view towards performing clustering and spatial prediction. A decomposition of the trace-variogram is proposed into amplitude and phase components and quantifies how spatial correlations between functional observations manifest in their respective amplitude and phase. This enables the generation of separate amplitude and phase clustering methods for spatial functional data and develops a novel spatial functional interpolant at unobserved locations based on combining separate amplitude and phase predictions. Through simulations and real data analyses, we demonstrate the advantages of the approach when compared to standard ones that ignore phase variation through more accurate predictions and more interpretable clustering results.

C1618: **Flexible functional data representation in higher dimensions using state space transformation**
Presenter: **Hiba Nassar**, Technical University of Denmark, Denmark

Functional data analysis (FDA) traditionally involves two steps: representing discrete observations as continuous functions and then applying functional methods to the represented data. The choice of the initial functional basis can significantly impact the outcomes of subsequent analyses. Recent research has demonstrated that data-driven spline bases can outperform predefined, rigid representations by using efficient knot placement through machine learning algorithms. However, extending these methods to higher-dimensional domains, such as images, presents challenges. The use of tensor-based spline spaces in such contexts requires knots to be placed on a lattice, which restricts the flexibility of knot placement, an essential aspect of effective modeling. A novel approach is introduced based on state space transformation, which accounts for the distribution of knots not through the direct construction of spline bases but by embedding knot selection into the underlying structure of the transformation. This method offers a more flexible and adaptable functional representation for higher-dimensional data. Preliminary results suggest this approach significantly enhances the ability to model complex data domains, particularly those with image-like structures.

C1408: **Flexible estimation of spatial covariance functions from multi-temporal DInSAR data**
Presenter: **Teresa Bortolotti**, Politecnico di Milano, Italy
Co-authors: Roberta Troilo, Alessandra Menafoglio, Simone Vantini

Sentinel-1 satellites offer extensive synthetic aperture radar data globally, revisiting locations every six days. Leveraging these data, differential interferometric processing techniques yield high-resolution ground displacement images that are accurate to millimeter precision. These insights into evolving ground conditions enable comprehensive monitoring of large areas prone to environmental hazards. Nonetheless, challenges arise when spatial units (e.g., water or vegetated areas) show a non-coherent variability across successive time instants, yielding missing values in single pixels or entire areas consistently missing along time. Although statistical reconstruction of missing data is possible through spatial interpolation (e.g., Kriging), it typically grounds on the second-order structure of the target field, which, besides being unknown, is typically characterized by strong non-stationarities. The challenge of estimating the spatial covariance operator is faced from time-evolving ground displacement images by developing a novel non-parametric methodology which grounds in the theory of functional data analysis. While the non-parametric approach guarantees flexibility to account for the non-stationarity of the field, a Laplacian-type regularization ensures continuity in the reconstructed operator. The methodology is showcased on ground displacement images collected to monitor the Phlegraean Fields, Italy, a region vulnerable to seismic and bradisismic activity.

C1409: **A new motif discovery based method for forecasting and imputation of functional data**
Presenter: **Jacopo Di Iorio**, Emory University, United States

Forecasting, involving the prediction of future values and/or the evolution of functional observations, has always been a major goal of functional data analysis. Given the increasing attention in the field of functional motif discovery, a method is proposed for functional forecasting that involves the identification of functional motifs, i.e., typical "shapes" or "patterns" recurring multiple times within a single curve and/or across misaligned portions of multiple curves. Portions characterized by the same motif are hypothesized to be more likely to evolve similarly; therefore, the identification of the last portion of a function before the prediction part as a portion can help in the forecasting exercise. Extensive diagnostics are performed to guide the user not only in tuning parameters but also in validating the aforementioned hypothesis, thus ensuring the applicability of the method. Method performance is assessed through simulations, and it is applied to a real-data case study. Similarly, the same methodology can be applied to the imputation of missing portions of a curve.

C1615: **Area-based epigraph and hypograph indices as a tool to detect outliers in functional data**
Presenter: **Belen Pulido Bravo**, Universidad Carlos III de Madrid, Spain

Co-authors: Rosa Lillo, Alba Franco-Pereira

The epigraph and hypograph indices offer alternative methods to functional depth when the goal is to order functions. Unlike functional depth, which provides a center-outward ranking, these indices produce top-to-bottom or bottom-to-top orderings. Modified versions of these indices, called ABEI and ABHI, are introduced based on the areas between curves. These new indices are considered here for detecting outliers in functional data. The primary advantage of ABEI and ABHI over their original formulations is their enhanced ability to isolate outliers more effectively. Furthermore, a novel procedure for outlier detection in functional data is proposed, utilizing ABEI and ABHI applied to the data as well as its first and second derivatives. This transforms the functional dataset into a multivariate one, enabling the application of established multivariate outlier detection techniques. The introduced approach is validated through extensive experimentation on both simulated and real-world datasets, demonstrating its competitive performance compared to existing methods in functional data analysis.

CO349 Room BH (S) 1.01 Lec. Theatre 1 CLIMATE ECONOMETRICS Chair: Helena Veiga

C0424: Adaptive now- and forecasting of global temperatures under smooth structural changes

Presenter: **Robinson Kruse-Becher**, FernUniversitat in Hagen, Germany

Accurate short-term now- and forecasting of global temperatures is an important issue and helpful for policy design and decision-making in the public and private sectors. A raw mixed-frequency data set is composed of weather stations around the globe (1920-2020). First, smooth variation is documented in average monthly and annual temperature series by applying a dynamic stochastic coefficient model. Second, adaptive cross-validated forecasting methods are used, which are robust to smooth changes of unknown form in the short run. Therein, recent and past observations are weighted in a mean squared error-optimal way. Overall, it turns out exponential smoothing methods (with bootstrap aggregation) often perform best. Third, by exploiting monthly data, a simple procedure is proposed to update annual nowcasts during a running calendar year and demonstrate its usefulness. Further, these findings are shown to be robust with respect to climate zones. Finally, now- and forecasting of climate volatility is investigated via a range-based measure and a quantile-based climate risk measure.

C1130: Anthropogenic warming increases the risk of major tropical cyclones in a nonstationary climate

Presenter: **Michael Wiper**, Universidad Carlos III de Madrid, Spain

Co-authors: Concepcion Ausin, Ali Sarhadi

The aim is to quantify the magnitude and frequency of maximum sustained wind speed of major tropical cyclones, which evolve under different nonstationary and warming climates at different locations in global basins. Bayesian spatiotemporal models are employed to quantify the zero wind probability and the mean and variance of non-zero wind at each site in a given time, and then a lognormal regression structure is fit using INLA. Based on the Laplace approximation, the INLA approach allows for much faster implementation than the standard MCMC methods. Both parametric and semi-parametric approaches are considered to estimate the development of the parameters in space and time and use these to approximate the future risk levels at each location.

C1415: Extreme temperatures and the profitability of large European firms

Presenter: **Gian Pietro Enzo Bellocca**, Universidad Carlos III de Madrid, Spain

Co-authors: Pilar Poncela, Esther Ruiz

The impact of temperature exposure is analyzed on the earnings per share of large European firms over the 21st century. Findings reveal that earnings are sensitive to extreme temperatures across a large proportion of sectors. Depending on the sector and quarter, exposure to extreme temperatures can have either a positive or negative impact on profitability. Analysis shows a greater percentage of sectors affected in Europe compared to the US, likely due to Europe's broader temperature variability from the northern Baltic to the southern Mediterranean regions. It is observed that most sectors experience effects during the milder seasons of spring and autumn, being positive in most cases. The lack of a clear negative effect of extreme temperatures on firms' profitability points out one of the reasons why it is so difficult to fight against climate change: while being harmful, it can be profitable. Additionally, a concerning trend is highlighted regarding a steady increase in European investments in sectors that are solely negatively impacted by extreme temperatures, which grew from around 16% in 2015 to over 23% in 2022.

C1461: Climate normals and anomalies

Presenter: **Tommaso Proietti**, University of Roma Tor Vergata, Italy

Co-authors: Alessandro Giovannelli

Studies of interannual climate variability rely on time series of temperatures, sea level, pressure, and rainfall anomalies. Anomalies are what residues after adjusting raw measurements for climate trends and seasonality. The adjustment is made by subtracting the decadal monthly or seasonal averages of the series computed over past reference periods consisting of three decades. Are the methods currently in use suitable? Are they able to capture climate warming trends and changing seasonality? Are traces of unadjusted trends seen in the anomaly series? Are the parametric and nonparametric methods adopted in other areas for similar tasks (seasonal adjustment) in econometrics a realistic alternative for climate time series? Alternative methods based on local polynomial and trigonometric regression can be devised to overcome some of the shortcomings of the methodology currently in use. The relevance of these alternative methods is discussed and evaluated with reference to the ENSO phenomenon.

C1383: Machine learning projection of climate and technology impacts on crops key to food security

Presenter: **Dan Li**, CSIRO, Australia

Co-authors: Vassili Kitsios, David Newth, Terry OKane

Climate change significantly threatens all economic sectors, with agriculture being especially vulnerable due to extreme conditions, rising temperatures, and shifting precipitation patterns. The impact on crop production will have critical implications for food security policies. Projecting the prospective impacts of climate change on crop production necessitates a comprehensive modelling system outlining crop responses to future conditions. A multivariate autoregressive econometric model is presented that incorporates a time-varying variable to reflect the diminishing impact of technology on crop yields. This model examines the interplay between technology, climate variables, and annual crop yield growth globally. By analyzing historical crop production and climate data from 1961 to 2018, the model outperforms traditional panel regression methods. Additionally, a novel machine learning climate model emulator enables efficient estimation of crop production under a multitude of carbon emission scenarios. Findings reveal that technological advancements may increasingly fail to counterbalance the adverse effects of climate change on wheat and rice production. This suggests that simplified climate damage functions in economic models can lead to significant inaccuracies in production estimates and flawed policy decisions.

CO142 Room BH (SE) 1.01 BAYESIAN STATISTICS Chair: Dimitri Konen

C1132: Valid uncertainty quantification for linear functionals in semi-parametric regression models

Presenter: **Gustav Romer**, University of Cambridge, Denmark

Inspired by Darcy's problem, the frequentist validity of Bayesian uncertainty quantification is addressed for irregular linear functionals in semi-parametric models. For a finite collection of linear functionals, a renormalized Bernstein-von-Mises theorem is proven to allow for posterior credible sets that are asymptotic confidence sets. This is demonstrated for a credible ellipsoid centered at the posterior mean and shaped by the posterior variance. A bound is provided on its diameter, and a Wald-type ellipsoid is introduced as an alternative. For a single linear function, results are obtained for the symmetric credible interval around the posterior mean. A general inverse problem is then analyzed, represented by a

Gaussian regression model, where the regression functions are parameterized by a non-linear forward map. A high-dimensional Gaussian prior is employed. Assuming the invertibility of the information matrix in high-dimensional approximation models, a renormalized Bernstein-von-Mises theorem is established for a finite collection of linear functionals. The conditions for the forward map induced by the partial differential equation, Darcy's problem, where irregular linear functionals naturally arise, are explicitly checked.

C1065: **Skewed Bernstein-von Mises theorem and skew-modal approximations**
Presenter: **Francesco Pozza**, Universita Bocconi, Italy

Gaussian deterministic approximations are routinely employed in Bayesian statistics to ease inference when the posterior distribution is intractable. Although these approximations are justified, in asymptotic regimes, by Bernstein-von Mises type results, in practice, the predicated Gaussian behavior may poorly represent the actual shape of the exact posterior, thereby affecting approximation accuracy. Motivated by these considerations, an improved class of closed-form and valid deterministic approximations of posterior distributions is derived, which arise from a novel treatment of a third-order version of the Laplace method yielding approximations within a tractable family of skew-symmetric distributions. Under general assumptions which allow to account for misspecified models, non-i.i.d. settings and various asymptotic regimes, this novel family of approximations is shown to have a total variation distance from the exact posterior whose rate of convergence improves by at least one order of magnitude the one achieved by the Gaussian from the classical Bernstein-von Mises theorem. The same improvement is also proved for polynomially bounded posterior functionals, and for a scalable strategy, approximate posterior marginals are derived. Through two real data applications, it is shown that the proposed approximations can be remarkably more accurate than their Gaussian counterparts.

C1100: **Semiparametric empirical Bayesian analysis of maxima and peaks over threshold**
Presenter: **Stefano Rizzelli**, University of Padova, Italy

Predicting future observations is the central goal of several statistical applications concerning extreme-value data. Under mild assumptions, extreme value theory justifies modeling linearly normalized sample maxima by max-stable distributions and rescaled excesses of a large threshold by Generalized Pareto distributions. The Bayesian paradigm offers a direct approach to forecasting and uncertainty quantification. Various Bayesian procedures have been proposed in recent years, though they typically disregard the asymptotic bias inherent in the use of extreme value models, incorporating no information on the norming sequences in the prior specifications for location and scale parameters. Some recently proposed empirical Bayes approaches have been reviewed to suitably address this point via data-dependent priors. The resulting asymptotic posterior concentration properties are illustrated, and their implications for estimation and prediction of future extreme observations are pinpointed.

C1123: **Output analysis for high-dimensional MCMC**
Presenter: **Ardjen Pengel**, University of Cambridge, United Kingdom

The widespread use of Markov chain Monte Carlo (MCMC) methods for high-dimensional applications has motivated research into the scalability of these algorithms with respect to the dimension of the problem. Despite this, numerous problems concerning output analysis in high-dimensional settings have remained unaddressed. Novel quantitative Gaussian approximation results are presented for a broad range of MCMC algorithms. Notably, the dependency of the obtained approximation errors is analysed on the dimension of both the target distribution and the feature space. It is demonstrated how these Gaussian approximations can be applied in output analysis. This includes determining the simulation effort required to guarantee Markov chain central limit theorems and consistent estimation of both the variance and the effective sample size in high-dimensional settings. Quantitative convergence bounds are given for termination criteria, and it is shown that the termination time of a wide class of MCMC algorithms scales polynomially in dimension while ensuring a desired level of precision. The results offer guidance to practitioners for obtaining appropriate standard errors and deciding the minimum simulation effort of MCMC algorithms in both multivariate and high-dimensional settings.

C1146: **Bayesian inference for killed reflected diffusions**
Presenter: **Fanny Seizilles**, University of Cambridge, United Kingdom
Co-authors: Richard Nickl

A setting of molecules diffusing independently in a multidimensional domain, following a killed reflected diffusion process. The Bayesian approach is investigated for nonparametric inference on the diffusion parameter D, from observations of the killing positions of the molecules.

CO046 Room BH (SE) 1.02 TOPICS IN BAYESIAN MODELING AND COMPUTATION Chair: Luca Maestrini

C0983: **Probabilistic activation functions and semiparametric mean field variational learning in Bayesian neural networks**
Presenter: **Mingwei Lin**, London School of Economics and Political Science, United Kingdom
Co-authors: Giulia Livieri, Luca Maestrini, Mauro Bernardi

Probabilistic representations for activation functions in Bayesian neural networks are proposed through the introduction of augmented variables. These models transcend traditional conjugate frameworks, which typically present intractability issues. To address these challenges, the semiparametric mean field variational approximations are implemented to manage the intractable posteriors in the parameter learning processes. Compared to Markov chain Monte Carlo (MCMC) methods, this approach maintains good approximation results and offers faster training processes. Additionally, a mixture of expert frameworks is introduced in variational learning, which further enhances prediction accuracy and holds substantial potential for reducing computational costs.

C0773: **A variational inference approach to variable selection for heteroskedastic regression models**
Presenter: **Giulia Livieri**, The London School of Economics and Political Science, United Kingdom
Co-authors: Mauro Bernardi, Luca Maestrini

Variable selection plays a key role in modern statistical research and learning. Major classes of variable selection approaches are implemented using Markov chain Monte Carlo methods. These methods may be computationally impractical for large-scale problems or complex models, and faster approximations are desirable or necessary. An approach to variable selection is developed for heteroscedastic regression models based upon semiparametric mean field variational Bayes. The proposed methodology is suitable for models having linear mean and exponential variance functions with prior specifications that induce sparse solutions on the regression coefficients. The use of classic mean field variational Bayes leads to the approximating densities having non-standard forms, and challenging numerical problems arise in the determination of the optimal approximation. Tractability is achieved by imposing a parametric assumption to the approximate marginal posterior densities of variance regression coefficients. The iterative optimization of the log-likelihood lower bound includes Newton-type steps with analytical derivative expressions for the parametric component of the approximation. This optimization strategy uses new results that solve recurrent issues of constrained optimization involving multivariate skew-normal variational approximations. Illustrations demonstrate the approach is computationally efficient and accurate in comparison to Markov chain Monte Carlo.

C0798: **Time-dependent stochastic block models with application to causes of death networks**
Presenter: **Cristian Castiglione**, Bocconi University, Italy

Discovering latent dependence structures over the nodes of a dynamic network is a difficult challenge which is of increasing importance in many applied fields. The interest is motivated by a demographic analysis of the complex interaction between underlying and multiple causes of death in a population observed on a fine age grid. To unveil non-trivial grouping structures between causes of death, a flexible stochastic block model is proposed for dynamic directed networks, which is able to learn asymmetric node partitions having common connectivity patterns. To flexibly account for the time evolution of the node grouping structure, a time-dependent random partition process is relied on, which permits the learning of sequences of partitions with a high level of persistence over time. The initial distribution of the sequence is then specified according to a Gibbs-type prior. This choice encompasses several routinely used prior distributions for Bayesian clustering, including fixed, random and infinite number of possible groups. An additional benefit of such a specification is to facilitate the inclusion of non-dynamic node-specific attributes in the model, which, in the application, permits the information of the partition sequence with external medical information on the causes of death. Alongside, an efficient algorithm to sample from the posterior distribution of the considered model is discussed.

C0952: **Fast Bayesian model selection algorithms for linear regression models**
Presenter: **Mauro Bernardi**, University of Padova, Italy
Co-authors: Manuela Cattelan, Claudio Busatto

The challenge of model selection in high-dimensional linear regression has traditionally been tackled by assuming hierarchical mixtures as prior distributions. To exclude irrelevant covariates, a spike component with Dirac probability mass at zero is introduced, leading to Bayesian selection procedures based on the marginal posterior distribution of various model configurations. Exploring the space of competing models involves computationally intensive simulation techniques. The issue of efficiently updating the variance-covariance matrix of the posterior distribution and the marginal posterior density following a change in the design matrix is addressed. Using thin QR factorization, novel algorithms are proposed to update the posterior variance-covariance matrix without storing and updating the Q matrix, resulting in significant computational savings. Furthermore, the focus is on evaluating the marginal posterior, a critical bottleneck in Bayesian model selection. The approach shows that computing the marginal posterior depends on the inverse of the R matrix. Thus, a methodology is developed to update both this inverse and the associated marginal posterior after modifying the design matrix. These methods eliminate the need for computationally intensive inversions of large matrices when evaluating the marginal posterior.

C0292: **Variational Bayesian Bow tie neural networks with shrinkage**
Presenter: **Alisa Sheinkman**, University of Edinburgh, United Kingdom
Co-authors: Sara Wade

Despite the dominant role of deep models in machine learning, limitations persist, including overconfident predictions, susceptibility to adversarial attacks, and underestimation of variability in predictions. The Bayesian paradigm provides a natural framework to overcome such issues and has become the gold standard for uncertainty estimation with deep models, also providing improved accuracy and a framework for tuning critical hyperparameters. However, exact Bayesian inference is challenging, typically involving variational algorithms that impose strong independence and distributional assumptions. Moreover, existing methods are sensitive to the architectural choice of the network. We address these issues by constructing a relaxed version of the standard feed-forward rectified neural network, and employing Polya-Gamma data augmentation tricks to render a conditionally linear and Gaussian model. Additionally, we use sparsity-promoting priors on the weights of the neural network for data-driven architectural design. To approximate the posterior, we derive a variational inference algorithm that avoids distributional assumptions and independence across layers and is a faster alternative to the usual Markov Chain Monte Carlo schemes.

CO189 Room BH (SE) 1.05 FINANCIAL MODELING: REGIME SWITCHING AND STATISTICAL LEARNING Chair: Christina Erlwein-Sayer

C0841: **Stock market responses to climate risks: Sectoral-level evidence from the U.S.**
Presenter: **Rogemar Mamon**, University of Western Ontario, Canada

Using a data set for all companies forming the S&P 500 index, the stock price responses to acute physical risks are investigated, chronic physical risks, and transition risks. The findings reveal that certain sectors are more vulnerable to climate risks, whereas others appear to be relatively unaffected. In addition, results show that listed firms with poor environmental performance scores are more exposed to climate risk, as indicated by their stock returns being negatively affected, compared to firms with higher environmental performance scores. This suggests that improving environmental performance may help companies to better cope with climate risks and improve their financial performances. The analysis provides evidence that the short-term systematic risk is more vulnerable to climate risk events, whereas effects on long-term systematic risk do not appear to be statistically significant. These findings indicate that investors and firms should pay particular attention to short-term systematic risk when considering the potential impact of climate risk on stock market performances.

C1012: **Modelling of financial time series with a regime-switching GARCH model including jumps**
Presenter: **Mai Phan**, University of Kaiserslautern-Landau & HTW Berlin, Germany
Co-authors: Joern Sass, Christina Erlwein-Sayer

Time series analysis helps identify underlying patterns and trends in financial data, enabling analysts to make informed predictions about future price movements. Several characteristics of financial time series make this analysis particularly challenging: volatility, non-stationary, heteroscedasticity, non-normal distribution and volatility clusters. Advanced time series models have been developed to address these challenges, such as a regime-switching GARCH model. This model allows for the identification of different regimes or states within financial time series, each with its own distinct volatility structure. By incorporating regime switching, the model can adapt to periods of high and low volatility and captures the dynamic behavior of financial markets. This adaptability makes the regime-switching GARCH model suitable for analyzing highly volatile financial time series like cryptocurrencies. Despite its advantages, the regime-switching GARCH model does not incorporate the occurrence of jumps in financial time series, whose frequencies can vary over time. To address this limitation, a regime-switching GARCH-jump model is proposed. This advanced model includes regime-specific jump frequencies and multiple states, along with GARCH processes tailored to each regime's conditional variance. This combination enhances the model's flexibility in representing the complex behavior of financial markets. This innovative approach is applied to model the daily log returns of Bitcoin.

C1014: **Sentiment analysis in an LSTM deep learning approach for forecasting volume in fractional trading**
Presenter: **Christina Erlwein-Sayer**, University of Applied Sciences HTW Berlin, Germany
Co-authors: Andrew Rosenswie, Alla Petukhina, Sakir Kepezkaya

Fractional trading has emerged as an opportunity for investors to invest in fractions of stocks. It gains popularity since trades in large companies and high-valued stocks can be realized even for small budgets and a bigger investment community. Risks of fractional trading for sellers and brokerages lie mostly in forecasting the number of fractional trades over a short period since those trades potentially need to be balanced out. News sentiment analysis and machine learning approaches are combined for time series modelling to find the best-suited models to forecast trading volume in fractional trading. A long-short-term memory neural network (LSTM) is built to forecast volume. The results are further investigated with measures of feature importance to find the most important explanatory variables. News sentiment analysis is combined with LSTM and most relevant features are discovered. The model and case studies focus on four major companies in different sectors to identify the most relevant features. The model is built in two steps: a cluster analysis finds company clusters with similar trading behaviors. In the second step, the volume is modelled through news-sentiment-based LSTM. Preliminary results show clear indications of the main explanatory variables, including news sentiment. Volume in fractional trading is modeled, and reliable features are found for forecasting volume. Findings show that the LSTM performance is mostly improved when news sentiment is added.

C1326: **Valuing real estate portfolios with machine learning using geospatial and macroeconomic data**
Presenter: **Sami Alkhoury**, HWR Berlin, Germany

The focus is on valuing real estate portfolios using publicly available data. Based on publicly registered real estate sales in the Ile-de-France region in the period 2014-2022, publicly available geospatial data (OpenStreetMap) and macroeconomic time series, various machine learning models are trained to forecast real estate sale prices. A particular focus is on the explainability and interpretability of the models. Techniques such as Shapley values and partial dependence plots help reveal the key factors driving property prices, providing clear and actionable insights.

C1631: **Bayesian Bandit portfolio: Customized Thompson sampling for investor preference**
Presenter: **Vlad Bolovaneanu**, Bucharest Academy of Economic Studies, Romania
Co-authors: Daniel Traian Pele

Investor preference contributes decisively to the structure of financial portfolios. Allocation techniques that favor a more conservative or risky approach is common practice. The advent of reinforcement learning (RL) for portfolio optimization created new opportunities for maximizing an arbitrary objective in the intricate market environment. Agents learn to navigate the market by maximizing a custom metric which rewards good decisions and punishes bad ones. At first, using only trial-and-error strategies, agents gradually improve over time. The multi-armed bandit problem, a well-known RL problem, has rarely been pursued in the portfolio optimization literature. A novel approach is proposed using Thompson sampling (TS), which deviates from current research on how sampling is performed. Whilst all works using TS so far have been interested in its standard form with a binomial distribution, sampling from the return distribution posterior is opted for. A tractable distribution is obtained via modeling tails as Pareto-distributed. The approach creates the possibility for an in-depth encoding of investor preference in the optimization metric and is competitive when compared to other well-known portfolios.

CO068 **Room BH (SE) 1.06** **RECENT DEVELOPMENTS IN THEORETICAL STATISTICS** **Chair: Yue Zhao**

C0191: **On counting communities and finding them**
Presenter: **Dana Yang**, Cornell University, United States

Random graph models with community structure have been studied extensively in the literature. For both the problems of detecting and recovering community structure, an interesting landscape of statistical and computational phase transitions has emerged. A natural unanswered question is: might it be possible to infer properties of the community structure (for instance, the number and sizes of communities) even in situations where actually finding those communities is believed to be computationally hard? It is shown that the answer is no. In particular, certain hypothesis testing problems between models with different community structures are considered, and it is shown (in the low-degree polynomial framework) that testing between two options is as hard as finding the communities.

C0381: **Statistical properties of a flexible regularized estimating equation formulation**
Presenter: **Yue Zhao**, University of York, United Kingdom

A central topic in high-dimensional statistics facing very many variables is regularisation, which seeks to retain only a few impactful variables in the estimated model. While regularised optimisation formulation has been well studied, much less has been said about the regularised estimating equation. The proposal is to formulate the zero root of a regularised estimating equation as the fixed point of a proximal operator specified by the regulariser. In this way, finding the zero root is translated into finding the fixed point of the proximal operator, for which many efficient algorithms exist. In addition, the said proximal operator is itself a simple convex problem and often admits closed-form solutions, even for more complex regularizers such as the non-convex (group) SCAD and MCP. The statistical properties of the solutions to the algorithm are presented, such as their variable selection consistency, in the non-asymptotic, high-dimensional regime. These solutions are shown to behave similarly to those from a comparable optimization formulation. Numerical studies also demonstrate the computational advantage of our algorithm.

C1519: **Optimal vintage factor analysis with deflation varimax**
Presenter: **Xin Bing**, University of Toronto, Canada

Vintage factor analysis is one important type of factor analysis that aims to first find a low-dimensional representation of the original data and then seek a rotation such that the rotated low-dimensional representation is scientifically meaningful. The most widely used vintage factor analysis is the principal component analysis (PCA), followed by the varimax rotation. Despite its popularity, little theoretical guarantee can be provided to date, mainly because varimax rotation requires solving a non-convex optimization over the set of orthogonal matrices. A deflation varimax procedure is proposed that solves each row of an orthogonal matrix sequentially. In addition to its net computational gain and flexibility, theoretical guarantees are fully established for the proposed procedure in a broader context. Adopting this new deflation varimax as the second step after PCA, this two-step procedure is further analyzed under a general class of factor models. The results show that it estimates the factor loading matrix in the minimax optimal rate when the signal-to-noise ratio (SNR) is moderate or large. In the low SNR regime, possible improvements are offered over using PCA and the deflation varimax when the additive noise under the factor model is structured. The modified procedure is shown to be minimax optimal in all SNR regimes.

C1433: **Adaptive ridge regression and fractional degrees of freedom**
Presenter: **Keith Knight**, University of Toronto, Canada

In classical regression model selection methods (such as forward and backwards selection, Mallows C_p, adjusted R^2, etc.), a predictor is either in the model or out of the model in the sense that its estimated parameter is assumed to be either a least squares estimate or 0, respectively. Shrinkage methods (such as ridge regression and the LASSO) can be viewed as relaxations of the classical model selection methods in the sense that they are able to shrink smaller least squares estimates close to 0 or exactly 0. An interesting property of ridge regression is that for given values of the ridge parameters, the fractional contribution (between 0 and 1) of each predictor can be defined in the model. It is considered to define the ridge parameters in terms of specified fractional contributions for each predictor.

C1322: **On the attainment of the Wasserstein-Cramer-Rao lower bound and the location scale family**
Presenter: **Hayato Nishimori**, The University of Tokyo, Japan
Co-authors: Takeru Matsuda

In information geometry, the Fisher information is regarded as a Riemannian metric that defines the local distance structure in the space of probability distributions. It also gives a lower bound on the variance of (unbiased) estimators in the Cramer-Rao inequality. On the other hand, it is recently reported that the Wasserstein distance, which is the optimal transportation cost between distributions, induces another Riemannian metric and an analogous inequality called the Wasserstein-Cramer-Rao inequality holds. The Wasserstein metric is obtained explicitly in statistical models on the real line by parametrizing the bijection that gives the push-forward measure instead of parametrizing the probability density directly. This parametrized bijection also gives the metric in multivariate models whose copulas do not depend on the parameter. In addition, the first and second-order moment estimators attain the Wasserstein-Cramer-Rao lower bound if and only if the statistical model is the location-scale family. Furthermore, if the statistical model is the location-scale family, estimators of the mean and variance asymptotically attain the lower bound.

CO172 **Room BH (S) 2.01** **FOUNDATIONS OF MACHINE LEARNING FOR ECONOMICS AND FINANCE** **Chair: Artem Prokhorov**

C0998: **Evaluating the statistical characteristics and analyzing the ML models of the limit order book**
Presenter: **Dragana Radojicic**, Faculty of Economics and Business, University of Belgrade, Belgrade, Serbia, Serbia

The purpose is to describe the dynamics and statistical properties of the limit order book. The goal is to analyze the informativeness of features extracted from the limit order book. The basis is on real market data from the Nasdaq Stock Market. Firstly, relevant features are extracted from the database, and then their informativeness is analyzed. A customized data processing reconstructor is utilized to extract essential structures from the limit order book. Since there is potential in both analyzing historical stock data and developing algorithms to interpret that data. Different feature extraction methods and performances of the neural network models based on different topologies are examined.

C1655: **Sparse sieve MLE**
Presenter: **Di Liu**, StataCorp, United States

The Neyman orthogonal estimator is derived in the Sieve MLE setting with parametric marginals and Bernstein copula.

C1135: **The price of independence in a model with unknown dependence**
Presenter: **Victor de la Pena**, Columbia University, United States

How much does it cost a decision-maker to base her payoff on interdependent, biased information sources? This question is relevant in economics, statistics and politics. When there are many information sources, their dependence may be unknown or uncertain, which creates multivariate ambiguity. One approach to answer one leading question involves decoupling inequalities from probability theory. New inequalities which hold for any type of dependence are presented. Applications to the problem of asset investment are provided.

C0273: **The mosaic permutation test: An exact and nonparametric goodness-of-fit test for factor models**
Presenter: **Asher Spector**, Stanford University, United States
Co-authors: Emmanuel Candes, Rina Foygel Barber, Trevor Hastie, Ronald Kahn

Financial firms often rely on fundamental factor models to explain correlations among asset returns and manage risk. Yet after major events, e.g., COVID-19, analysts may reassess whether existing risk models continue to fit well: specifically, after accounting for the factor exposures, are the residuals of the asset returns independent? With this motivation, the mosaic permutation test is introduced, a nonparametric goodness-of-fit test for preexisting factor models. The method can leverage nearly any machine learning technique to detect model violations while provably controlling the false positive rate, i.e., the probability of rejecting a well-fitting model, without making asymptotic approximations or parametric assumptions. This property helps prevent analysts from unnecessarily rebuilding accurate models, which can waste resources and increase risk. To illustrate the methodology, the mosaic permutation test is applied to the BlackRock fundamental equity risk (BFRE) model. Although the BFRE model generally explains the most significant correlations among assets, evidence of unexplained correlations is found among certain real estate stocks, and it is shown that adding new factors improves model fit. The methods in the Python package mosaicperm are implemented.

C1654: **Improved semi-parametric bounds for tail probability and expected loss: Theory and applications**
Presenter: **Artem Prokhorov**, University of Sydney, Australia

Many management decisions involve accumulated random realizations for which the expected value and variance are assumed to be known. The tail behavior of such quantities is revisited when individual realizations are independent, and new sharper bounds are developed on the tail probability and expected linear loss. The underlying distribution is semi-parametric in the sense that it remains unrestricted other than the assumed mean and variance. The bounds complement well-established results in the literature, including those based on aggregation, which often fail to take full account of independence and use less elegant proofs. New insights include proof that in the non-identical case, the distributions attaining the bounds have the equal range property and that the impact of each random variable on the expected value of the sum can be isolated using an extension of the Korkine identity. It is shown that the new bounds not only complement the extant results but also open up abundant practical applications, including improved pricing of product bundles, more precise option pricing, more efficient insurance design, and better inventory management. For example, a new solution to the optimal bundling problem is established, yielding a 17% uplift in per-bundle profits, and a new solution to the inventory problem, yielding a 5.6% cost reduction for a model with 20 retailers.

CO234 Room BH (S) 2.02 ADVANCED STATISTICAL TECHNIQUES: FROM RISK TO AI AND BEYOND Chair: Silvia Montagna

C1666: **A copula-based data augmentation strategy for the sensitivity analysis of extreme operational losses**
Presenter: **Amir Khorrami Chokami**, University of Cagliari, Italy
Co-authors: Giovanni Rabitti

The aim is to assess the importance of macroeconomic and financial variables for UniCredit Bank's operational losses. To achieve this, the Shapley effects is considered a variance-based measure of importance. However, the small number of observations of extreme losses makes the estimation of the Shapley effects challenging. To address this issue, augmenting the sample of extreme observations is proposed, using vine copulas and calculating the Shapley effects on the augmented sample. The effectiveness of this procedure is supported by a numerical simulation. Findings obtained with the methodology applied to the UniCredit Bank data show its usefulness for the risk management of operational losses.

C1649: **Identifying microbiome communities and enterotypes using a novel mixed-membership model**
Presenter: **Roberto Ascari**, University of Milano-Bicocca, Italy
Co-authors: Alice Giampino, Sonia Migliorati

Understanding how the human gut microbiome affects host health is challenging due to the wide interindividual variability, sparsity, and high dimensionality of microbiome data. Recently, mixed-membership models have been applied to these data to detect latent communities of bacterial taxa that are expected to co-occur. The most widely used mixed-membership model is the latent Dirichlet allocation (LDA). However, LDA is limited by the rigidity of the Dirichlet distribution imposed on the community proportions, which hinders its ability to model dependencies and account for overdispersion. To address this limitation, a generalization of LDA that introduces greater flexibility into the covariance matrix is proposed by incorporating the flexible Dirichlet (FD). In addition to identifying communities, the new model enables the detection of enterotypes, i.e., clusters of samples with similar microbe composition. A computationally efficient collapsed Gibbs sampler is proposed that exploits the conjugacy of the FD distribution with respect to the multinomial model. A simulation study demonstrates the model's ability to recover the true parameter values and the correct number of communities. Moreover, an application to the COMBO dataset highlights its effectiveness in detecting biologically significant communities and enterotypes, underscoring the new model as a definite improvement over LDA.

C1639: **Learning the distribution map in reverse causal performative prediction**
Presenter: **Daniele Bracale**, University of Michigan, United States
Co-authors: Yuekai Sun, Moulinath Banerjee, Subha Maity

In numerous predictive scenarios, the predictive model affects the sampling distribution; for example, job applicants often meticulously craft their resumes to navigate through screening systems. Such shifts in distribution are particularly prevalent in the realm of social computing, yet the strategies to learn these shifts from data remain remarkably limited. Inspired by a microeconomic model that adeptly characterizes agents' behavior within labor markets, a novel approach is introduced to learn the distribution shift. The method is predicated on a reverse causal model, wherein the predictive model instigates a distribution shift exclusively through a finite set of agents' actions. Within this framework, a microfoundation model is employed for the agents' actions, and a statistically justified methodology is developed to learn the distribution shift map, which is demonstrated to be effective in minimizing the performative prediction risk.

C1671: **Enhancing embedding models through specialized fine-tuning in the banking sector**
Presenter: **Claudia Berloco**, Intesa Sanpaolo, Italy
Co-authors: Enrico Capuano

The widespread development of generative AI and natural language processing has led to the adoption of embedding models in various applications, including question-answering systems. These systems rely on the representation of words and sentences through embeddings to retrieve relevant information before feeding the large language models (LLMs). However, open-source pre-trained multipurpose embedding models may not capture specific nuances in certain contexts, such as the banking sector. Fine-tuning pre-trained embedding models on a dedicated dataset is investigated to improve their performance in specific contexts. A proprietary dataset is constructed from banking sector documents and divided into training and test sets. Various pre-trained open-source multipurpose embedding models are evaluated on the test set and fine-tuned on the training set. Performance is also assessed within a retrieval augmented generation (RAG) pipeline. The results are compared to those of the original multipurpose models to determine the impact of fine-tuning on sentence comprehension and retrieval. Analysis of the fine-tuned models' performance provides insight into tailoring embedding models to meet the unique needs of various industries and applications.

C1672: **Displaying the performance-consumption tradeoff for aware and sustainable AI**
Presenter: **Enrico Capuano**, Politecnico di Torino, Italy
Co-authors: Claudia Berloco

The rapidly evolving field of generative AI has witnessed a proliferation of AI models, often prioritizing performance metrics at the expense of energy consumption. As the academic community and civil society increasingly focus on environmental, social, and governance (ESG) issues, concerns about the environmental impact of resource-intensive computations in natural language processing (NLP) are growing. Large language models (LLMs) require substantial computational power during both training and deployment stages. In contrast, machine learning and statistical learning models are less computationally intensive, albeit capable of accomplishing different tasks. The aim is to raise awareness among developers and users about the energy consumption associated with different models by comparing their performance on a common text classification task. The trade-offs between prediction accuracy and power consumption are highlighted by juxtaposing performance metrics with energy consumption. The proposed methodology promotes responsible and sustainable AI development practices, enabling researchers and practitioners to make informed decisions that balance accuracy, speed, and energy efficiency. By promoting transparency and accountability in AI model development, the contribution is to the broader discourse on sustainability and responsible innovation in artificial intelligence.

CO311 Room BH (S) 2.03 MODELING AND MEASURING MULTIVARIATE VOLATILITY AND RISK Chair: Ilya Archakov

C0786: **Cardinality constraints meet large-scale portfolio**
Presenter: **Yuan Chen**, University of Vienna (VGSF), Austria
Co-authors: Nikolaus Hautsch, Bo Peng, Immanuel Bomze

In financial econometrics, the focus is often on improving covariance matrix estimations rather than addressing optimization problems with constraints for better portfolio management. It is argued that combining these advanced estimation methods with optimization that includes specific limits, like cardinality constraints, enhances decision-making and investment strategies. Cardinality constraints limit the number of assets in a portfolio, potentially making simpler estimators like the sample covariance sufficient for investment decisions, especially when dealing with large dimensions that typically introduce significant estimation errors affecting portfolio performance. The issue of managing portfolios is also addressed when there are fewer data points than assets, leading to non-invertible, noisy covariance matrices. Cardinality constraints simplify this challenge, making it possible to aim for a global minimum variance portfolio despite these limitations. Empirically, it is found that smaller portfolios, constrained by cardinality to include only a subset of available assets, can achieve diversification similar to market portfolios while reducing transaction costs and simplifying analysis. This suggests focusing on smaller, strategically selected portfolios could offer investors efficient and cost-effective outcomes.

C0679: **Dynamic factor model for realized covariance matrices**
Presenter: **Jasper Rennspies**, University of Freiburg, Germany
Co-authors: Ilya Archakov, Roxana Halbleib

A dynamic factor model is developed for realized covariance matrices. The log-transformation is used, as proposed in a prior study, to decompose realized covariance matrices into realized volatilities and transformed realized correlations. The panel of the resulting series is modelled by aggregating AR(1)-factors to capture persistence in a parsimonious way. A standard Kalman filter is used together with maximum likelihood to extract the latent factors and estimate the model parameters. The Kalman Filter setting is extended to allow for GARCH-type effects in the factors. The aggregation approach allows for an interpretation of the persistence in the factors that is data-driven. An empirical analysis of 29 liquid stocks on the NYSE is performed. Descriptive statistics of the 406 series of transformed realized correlations are reported. An in-sample analysis is performed, in which we inspect the autocorrelation in the residuals, and an out-of-sample analysis is performed for various forecast horizons.

C1168: **Forecast relative error decomposition**
Presenter: **Quinlan Lee**, University of Toronto, Canada
Co-authors: Christian Gourieroux

A class of relative error decomposition measures is introduced that are well-suited for the analysis of shocks in nonlinear dynamic models. They include the forecast relative error decomposition (FRED), forecast error Kullback decomposition (FEKD) and forecast error Laplace decomposition (FELD). These measures are favorable over the traditional forecast error variance decomposition (FEVD) because they account for nonlinear dependence in both a serial and cross-sectional sense. This is illustrated by applications to dynamic models for qualitative data, count data, stochastic volatility and cyber risk.

C1308: **Online forecasting of unbalanced implied volatility surfaces**
Presenter: **Arnaud Dufays**, EDHEC Business school, France
Co-authors: Jeroen Rombouts, Kris Jacobs

The daily option implied volatility surface is difficult to forecast with standard time series models because of its time-varying granularity. Approaches using option pricing models face a time-consuming estimation problem because realistic models require multiple latent factors. To address these challenges, a sequential surface forecasting approach is proposed that involves daily fitting combined with a dynamic model based on the parameter estimates as summary statistics of the previously implied volatility surfaces. The method works with any surface fitting method, such as option pricing processes, nonparametric methods, and machine learning models. In the empirical application to forecasting S&P 500 implied volatility surfaces, it is found that nonparametric and machine learning approaches typically outperform advanced option pricing models. The dynamic model with a particular case called the Surface HAR model, is shown to generally lead to significant forecast improvements for all models.

C1397: **Estimation risk for systemic risk measures driven by semi-parametric models**
Presenter: **Jeremy Leymarie**, ESC Clermont Business School, France

The purpose is to provide an analytical method to quantify the estimation risk contained in the systemic risk measures used to identify the financial institutions that contribute the most to the overall risk in the financial system. Estimation of the marginal expected shortfall (MES) and the delta conditional value-at-risk (Delta CoVaR) are investigated, when the firm and market returns follow a semi-parametric bivariate dynamic model. A two-step filtered historical simulation method, which is assumption-free on the distribution of the innovations, is proposed to estimate these quantities. We develop the asymptotic theory for the MES and the Delta CoVaR. The proposed method is evaluated by simulation and proved valid. Then, the method is applied to a panel of U.S. financial institutions to test whether the systemic risk contribution of a financial entity is significant on a single day, e.g. on September 15th 2008, given the information set prior to this date.

CO259 **Room BH (SE) 2.09** **EMPIRICAL MACRO** **Chair: Michael Owyang**

C0491: **Lost in aggregation: European, country, sectoral, and regional factors driving the GVA fluctuations in Europe**
Presenter: **Aikaterini Karadimitropoulou**, University of Piraeus, Greece
Co-authors: Krzysztof Beck

Ongoing monetary integration in Europe requires close monitoring of the degree of business cycle (BC) synchronization to assess the effectiveness of the common monetary policy. A Bayesian dynamic latent factor model is estimated using disaggregated real gross value added (GVA) data that includes four different types of factors: European, country, sectoral and regional. A richer factor structure reduces the variance attributed to the European factor (9%), while country, sectoral, and regional factors account for 26%, 21%, and 27%, respectively. Sectoral factors are the main drivers of international BC. Sub-periods analysis shows that the share of variance explained by the European factor increased modestly, while the share explained by the sectoral factor increased significantly at the expense of the country factor. The results support the European Commission's view on the synchronization of BC in the monetary union. Next, using Bayesian model averaging, the country factor is prevalent in regions characterized by the highest degree of specialization and belonging to countries with the least developed financial markets and most volatile exchange rates. The sectoral factor is most prevalent in regions that are part of countries with the most developed financial markets. Finally, the importance of commodity, monetary, fiscal, productivity, and terms of trade factors is examined for the European and country factors using factor-augmented vector autoregression models.

C0432: **Assessing liquidity constraints: Does credit matter for the permanent income hypothesis**
Presenter: **Neville Francis**, UNC Chapel Hill, United States

Recently, the validity of the permanent income hypothesis (PIH), a fundamental consumer behavior theory first proposed by an early study, has been questioned. A recent study found that when credit cards were issued to individuals who didn't previously have them, those with the highest savings experienced the largest change in consumption spending. This finding is out of line with the PIH's predictions: the PIH would predict that individuals with the least wealth would find the most relief from this new credit line and should react more. It is argued that the PIH can be rescued, and a model in which the above counterintuitive results are consistent with the PIH is proposed. The argument rests on the steadystate to which we make comparisons. In particular, it is argued that credit margins are important for model outcomes. The incentive to use cash versus credit hinges on the money already in hand. Recent studies have focused on the extensive margins and reached conclusions that question the theory's validity. However, the PIH theory may persist if a model is considered to expand rather than introduce credit. The aim is to propose a model in which, in comparison to a no-credit equilibrium, the extensive credit margin appears to invalidate the PIH, but in the presence of a steady state with credit, the intensive credit margin the PIH holds.

C0537: **The relative importance of news and knowledge**
Presenter: **Amy Guisinger**, Lafayette College, United States
Co-authors: Michael Owyang, Michael McCracken

Previous studies show the Fed has a forecast advantage over the private sector in terms of inflation. The purpose is to evaluate how much of this advantage results from the Fed's knowledge of future monetary policy and how much results from the difference in the two forecasts' timing. The previous method of equalizing the Fed's and private sectors' information sets is expanded on, and it is found that Fed forecasts do not encompass those of the private sector when the two have similar information sets.

C0808: **Expectations vs data news on nowcasting US GDP**
Presenter: **Ana Galvao**, Bloomberg Economics, United Kingdom

Nowcasting models typically rely on the predictive content of a set of indicators to anticipate the first estimate of key economic variables such as GDP growth and employment. The accuracy of these models is frequently not significantly better than consensus expectations surveyed just before the release. The predictive content of monthly indicators is compared to their consensus expectations to nowcast US GDP growth. Optimal weights are estimated for both types of predictors using a Bayesian machine learning method applied to a mixed-data sampling regression (MIDAS). Empirical results support the usefulness of surveyed expectations in forecasting GDP growth in real time.

C0812: **International stock co-movements and time-varying risks**
Presenter: **Michael Owyang**, Federal Reserve Bank of St Louis, United States
Co-authors: Laura Coroneo, Laura Jackson Young

International stock return comovements of country-industry portfolios are examined. The model allowed comovements to be driven by a global and a cluster factor, with the cluster membership endogenously determined. It is found that country-industry portfolios tend to cluster mainly within geographical areas. The cluster compositions substantially changed over time, with the emergence of clusters among European countries from the early 2000s. The cluster component was the main driver of country-industry portfolio returns for most of the sample, except from the mid-2000s to the mid-2010s, when the global component had a more prominent role. To this model, asymmetric risk is accounted for by adding time-varying skewness to the factors. Because the risk of interest is assumed to be relatively persistent, the skewness is modeled as Markov-switching. It is found that using a standard normal distribution always results in larger volatilities, no matter how the clusters are specified; using the skew-normal distribution results in lower volatilities, with a marginal improvement in the definitions of the endogenous clusters. Changes in volatility and factor loadings for all the specifications are also accounted for.

CO161 **Room BH (SE) 2.10** **STATISTICAL MODELS AND METHODS FOR EDUCATION II** **Chair: Antonella D Agostino**

C0563: **A mediation analysis approach for gender pay gap in STEM: The university of Palermo case**
Presenter: **Giovanni Boscaino**, University of Palermo, Italy
Co-authors: Martina Vittorietti, Ornella Giambalvo

Italy's strategic sustainability plan for 2024-2026 aims to reduce the wage disparity between men and women to a 1% difference. Addressing the gender pay gap (GPG) is crucial to promoting equity, improving economic efficiency and ensuring sustainable and inclusive development. Italy's GPG is around 5%, with an annual men vs women disparity of almost 8,000 between genders in the private sector in 2022. Additionally, a significant factor in the GPG is the underrepresentation of women in STEM fields, which offer higher wages. The hypothesis is that participation in STEM fields mediates the relationship between gender and wages. Using data from the University of Palermo graduates, the gender effect on wages is decomposed into indirect and direct effects through STEM participation. In addition, the propensity score approach is used to relate the study to an experimental one and a quantile regression model is adopted to investigate better the effect of the variables on the income distribution.

C0594: **Identifying exchange patterns in Erasmus+ mobility**
Presenter: **Kristijan Breznik**, International School for Social and Business Studies, Slovenia
Co-authors: Giarncarlo Ragozini, Marialuisa Restaino, Maria Prosperina Vitale

The international flow of students across countries has increased over the years. It has become important for higher education institutions to analyze these flows and understand the main characteristics of student mobility trajectories involved in international exchange programs. Data from the European Union open data portal is used, and the focus is on identifying highly connected countries and institutions. Specifically, using a network analytic approach, the aim is to determine whether cooperation patterns exist among European countries and higher education institutions, with some of them potentially considered as leading actors of international exchange. A significant part of the analysis revolves around the concept of the rich-club phenomenon. This concept refers to the tendency of highly connected nodes within a network to form tightly connected groups, which in the context translates to certain institutions becoming central hubs of student mobility. By testing for the rich-club phenomenon in the network of institutions involved in Erasmus and Erasmus+ mobilities, the purpose is to uncover whether a select group of institutions dominates the international student exchange community. Findings on the presence and implications of rich club structures can provide valuable insights into the dynamics of international cooperation in higher education.

C0597: **Exploring factors affecting gender gap in university student performance**
Presenter: **Marialuisa Restaino**, University of Salerno, Italy
Co-authors: Michele La Rocca, Marcella Niglio, Maria Prosperina Vitale

In recent years, exploring the determinants that may influence students' achievement in higher education has received much attention. Empirical studies have found that the most important factors affecting student performance are socio-demographic characteristics and attitudes, family context, results obtained at high school, social interactions among peers, and geographical areas. The contribution investigates the differences in students' performances attending a degree program in science, technology, engineering, and mathematics (STEM) by using statistical models to mainly capture the presence of a gender gap. University performance is measured by the number of ECTS credits students earn during the first year, which represents an important stage of their career path. Starting from a cohort of students enrolled at STEM degree programs in Italian universities, the main purposes are to estimate the probability of getting at least a certain number of credits at the end of the first year as well as to capture if there are any differences in performance between male and female investigating the factors affecting these results. Secondary data obtained from the National Student Archive are used to perform statistical models that capture the effect of contextual and individual factors on student performance.

C0598: **A discrete-time mover and stayer model with time-varying covariates for studying Italian student mobility**
Presenter: **Martina Vittorietti**, Delft University of Technology, Netherlands
Co-authors: Eni Musta

The mover-stayer model, also known as the "cure model", is a model to study social change over time in a heterogeneous population. It extends the first-order Markov process by recognizing a subgroup of individuals called "stayers," who will not experience the target event. Traditionally, the probability of being a stayer is the same for all individuals and corresponds to the proportion of stayers in any given state. However, this probability can be influenced by fixed-time covariates and time-varying covariates. The inclusion of the latter is not trivial. When subjects are assessed periodically over a certain period, panel data is dealt with. Panel data are common in the education field. A new dynamic version of the discrete mover-stayer model is presented using multinomial logistic regression with time-varying covariates, specifically focusing on panel data on the mobility of Italian Master's students. Factors such as the students' undergraduate degrees and the rankings of their universities are considered time-varying covariates in modeling their probability of moving; sex and age at enrollment as fixed-time covariates. The model employs a maximum likelihood estimation approach, integrating both the multinomial responses and the stayer status. Its identifiability is evaluated, simulation studies are conducted, and its performance is compared against established models in the literature.

C1416: **Q-matrix estimation in cognitive diagnostic models by using overlapping clustering**
Presenter: **Atsuro Kimoto**, Doshisha University, Japan
Co-authors: Jun Tsuchida, Hiroshi Yadohisa

Cognitive diagnostic models (CDMs) are educational measurement models that assess individuals' mastery of certain cognitive skills (hereafter, attributes) from their responses. A key objective of CDMs is to extract the attribute profile that represents each individual's attribute mastery status. To extract the attribute profile, the Q-matrix, which associates each test item with the attributes, should be developed before applying the CDMs. While domain experts generally develop the Q-matrix, it can be mis-specified, potentially leading to incorrect interpretations. Various data-driven Q-matrix estimation methods have been proposed to address this issue. However, estimating the Q-matrix appropriately is difficult because the elements of the Q-matrix are binary variables, and multiple attributes can be associated with a single test item. To address this difficulty, a novel Q-matrix estimation method is proposed based on the elements of the Q-matrix as a membership indicator of overlapping clustering. Using the overlapping clustering algorithm is expected to improve both accuracy and computational efficiency. Moreover, the proposed method can be applied to several types of CDMs. The results of numerical experiments show that the proposed method is superior in accuracy and computational efficiency compared to existing methods.

CO304 Room BH (SE) 2.12 STATISTICS IN GEOSCIENCES Chair: Radomyra Shevchenko

C1227: **Statistics and structures: Examples from oceanography**
Presenter: **Jonathan Lilly**, Planetary Science Institute, United States

In seeking to extract information from large oceanographic datasets, advances have come when a model for the structure of a particular type of phenomenon is combined with relevant statistical considerations. Three examples are given: (i) isolated, bumplike anomalies in sea surface height profiles observed by satellites, associated with coherent oceanic vortices; (ii) quasi-oscillatory features in data from freely-drifting instruments, associated with fluid trapping within these vortices; and (iii) random motions of large-scale fluid turbulence observed by freely-drifting instruments. In the first two cases, a structural model is created for isolated events and for modulated oscillations, respectively; both types of features are then detected using the continuous wavelet transform, and statistical significance is assessed by comparing two key feature properties with the distributions expected under a null hypothesis of unstructured red noise. The 2D survival function or complementary cumulative distribution function emerges as a useful statistical quantity. In the third example, a three-parameter stochastic model essentially mimicking real-world trajectory behavior is created using the Matern process, shown to be rightly thought of as a damped version of fractional Brownian motion. A general lesson from these efforts is the central importance of a suitable structural model as a basis for framing statistical questions.

C1306: **Machine learning and spatiotemporal statistics in the ocean: Fusing data sources and making nonlinear predictions**
Presenter: **Adam Sykulski**, Imperial College London, United Kingdom

The ocean is observed through a variety of means, including from satellites (remotely) and from instruments deployed at sea (in-situ). Sometimes, these measurements agree, albeit with different observational noise and sampling characteristics, but sometimes, they measure fundamentally different but related quantities. Reconciling these data sources is therefore desirable but poses a significant challenge, and how best to do it depends on the task at hand. Two examples of how machine learning and spatiotemporal statistics can be used are presented to fuse heterogeneous data sources in a nonlinear sense to make better predictions. The first example attempts to predict ocean surface velocities using satellite data only but trained using floating instruments called drifters. A probabilistic prediction framework that is implemented using a novel multivariate natural gradient boosting approach is used. The second example fuses satellite and drifter data to predict the abundance of Antarctic Krill in the Southern Ocean. As the data is "zero-inflated" in that krill tend to swarm and can otherwise be absent from large regions of the ocean, a Hurdle Gamma model is employed to first model presence-absence and then predict the abundance, all the time using many covariates from satellites and instruments at sea to improve predictions in a nonlinear sense.

C1137: **Scale interactions in the ocean: Designing and analyzing stochastic turbulence closures in complex ocean models**
Presenter: **Stephan Juricke**, GEOMAR Helmholtz Centre for Ocean Research Kiel and Constructor University Bremen, Germany

Ocean turbulence encompasses complex dynamics, ranging from micro- to planetary scales. Kinetic and potential energy are continuously converted and transferred in time and space. Explicitly modelling these chaotic interactions in global numerical ocean simulations is not feasible due to limited computational resources. One has to resort to limited grid resolution for numerical discretizations, ultimately neglecting small-scale processes that are crucial for correctly capturing the climate system's energy cycle. To still mimic underlying physical interactions, one applies so-called turbulence parameterizations that try to capture some of the statistical properties of unresolved small-scale processes relevant to the resolved, large-scale flow. Several deterministic and stochastic parameterization schemes are presented for so-called oceanic mesoscale turbulence, which is driven by physical instabilities on scales between 1-100km. The schemes are designed to inject energy that has been excessively dissipated in the simulation due to a misrepresentation of relevant energy transfers. Statistical scale analysis is used better to understand energy transfers with and without the parameterizations. Higher resolution simulations where mesoscale instabilities are largely resolved are used to constrain better and improve turbulence closures and their statistical properties. These new developments improve representations of oceanic flow, reducing systematic model biases.

C1658: **Generative modelling and stochastic parametrizations for a rotating shallow water system**
Presenter: **Oana Lang**, Imperial College London, United Kingdom
Co-authors: Dan Crisan, Alexander Lobbe

The stochastic parametrization of small-scale processes is essential in the estimation of uncertainty when trying to represent systematic model errors which arise from small-scale fluctuations (e.g. weather and climate predictions). In a stochastic partial differential model, the noise can be calibrated in a way that is consistent with such subgrid-scale parametrizations using a principal component analysis (PCA). The focus is on the explanation of how the PCA technique can actually be replaced by a generative diffusion model technique - this allows avoiding the imposition of additional constraints on the increments. This methodology is applied to a stochastic rotating shallow water model, using the elevation variable of the model as input data.

C1134: **Multi-scale CUSUM tests for time dependent spherical random fields**
Presenter: **Anna Vidotto**, University of Naples Federico II, Italy
Co-authors: Alessia Caponera, Domenico Marinucci

The purpose is to investigate the asymptotic behavior of structural break tests in the harmonic domain for time-dependent spherical random fields. In particular, a functional central limit theorem result is proven for the fluctuations over time of the sample spherical harmonic coefficients under the null of isotropy and stationarity; furthermore, consistency of the corresponding CUSUM test is proven under a broad range of alternatives, including deterministic trend, abrupt change, and a nontrivial power alternative. The results are then applied to NCEP data on global temperature: the estimates suggest that climate change does not simply affect global average temperatures but also the nature of spatial fluctuations at different scales.

CC504 Room S0.03 STATISTICAL METHODS AND APPLICATIONS I Chair: Catia Scricciolo

C1705: **Propensity score weighting with complex survey data: Best practice**
Presenter: **Guangyu Tong**, Yale University, United States
Co-authors: Yukang Zeng, Fan Li

A unified framework is introduced for integrating survey weights into propensity score weighting methods for causal inference with complex survey data. We incorporate survey weights into both propensity score estimation and outcome modeling under the balancing weights framework, establishing the asymptotic normality of survey-weighted estimators and extending the minimum variance property of overlap weights to survey settings. Furthermore, we develop three augmented estimators: moment, clever covariate, and weighted regression, each specifically adapted for survey weights to enhance robustness and efficiency. We also provide robust empirical closed-form sandwich variance estimators to ensure reliable variance estimation in survey settings. Through multistage sampling simulations designed to mimic real-world data collection scenarios, we demonstrate that balancing weights based on survey-weighted propensity scores can achieve effective population balance and provide consistent estimates for key population-level causal estimands, including PATE, PATT, and PATO. Our findings indicate that the augmented estimators effectively reduce bias and enhance efficiency, with the weighted regression estimator showing particular robustness across varying levels of covariate overlap and in the presence of model misspecification.

C1709: **Statistical analysis of data from supersaturated split-plot experiments**
Presenter: **Zhuowei Liang**, Kings College London, United Kingdom
Co-authors: Kalliopi Mylona

The supersaturated split-plot designs (SSPDs) are screening designs that have more potentially active factors than the number of experimental units under restricted randomisation due to the presence of hard-to-change factors. SSPDs reduce the experimental cost drastically, however, conventional statistical analysis methods are not applicable due to the large p small n feature and the random effect induced by restricted randomisation makes the data analysis even more challenging. We will present a Bayesian method for analysing data from SSPD experiments.

C1707: **A Bayesian approach to discrimination-free insurance pricing with variational inference**
Presenter: **Shuang Zhou**, Arizona State University, United States
Co-authors: Lydia Gabric, Kenneth Zhou

In recent years, many jurisdictions have implemented anti-discrimination regulations that require protected information to be excluded from insurance pricing calculations. To adapt to anti-discrimination regulations while maintaining accurate pricing outcomes, recent studies have sought to develop discrimination-free pricing methods through probabilistic inference. However, most of the existing estimation processes require individual-level discriminatory data, which are often prohibited due to regulatory constraints. We propose a novel Bayesian discrimination-free pricing method that no longer requires individual discriminatory information. To achieve this, we consider a Bayesian finite mixture model that treats the discriminatory variables as unknown latent variables with a hierarchical prior distribution to represent the assumed discriminatory relations. Through such model, the indirect discrimination can be inferred via the posterior distribution. Based on the hierarchical model, we propose a novel implementation of the variational inference to construct the discrimination-free pricing family with a mean-field family. Additional techniques such as the importance sampling are also used to obtain accurate prices. Supported by a simulation study and an empirical analysis based on real insurance data, our Bayesian approach is capable of inferring the hidden relationship between variables and consequently producing unbiased discrimination-free pricing results.

C1712: **Robustness in weighted networks**
Presenter: **Luisa Cutillo**, University of Leeds, United Kingdom
Co-authors: Dario Righelli, Valeria Policastro, Annamaria Carissimo

In network analysis, numerous community detection algorithms have been developed, yet their statistical validation remains underexplored. The R package ROBIN introduced a method for testing the robustness of unweighted networks using a configuration model as a null hypothesis to determine if detected communities result from random edge placement. We extend this approach to weighted networks and propose a new machine learning-based method for robustness analysis. The goal is to validate the robustness of community detection algorithms and compare their performance under various perturbation strategies. For weighted graphs, we employ the weighted configuration model (WCM), which preserves node strength sequences while randomizing edge positions and weights. This preserves the heterogeneity of node strengths, allowing for a statistical test of community structures. Our perturbation strategy rewires a percentage of the network's edges while maintaining the degree and weight distributions. This controlled perturbation ensures that key network properties are preserved. We tested this method using the LFR benchmark model, which simulates networks with power-law distributions for degree and community sizes. Various configurations were explored to evaluate algorithm robustness. Results demonstrate significant differences in algorithm stability, providing insights into which methods are more reliable for detecting communities in weighted networks under perturbation.

C1120: **Independent reads of Poisson flows**
Presenter: **Fatemah Alqallaf**, Kuwait University, Kuwait

The independent readings of Poisson flows are studied by two different sources. The primary concern of the model is to determine a bivariate random vector with compound Poisson distribution components generated by the flow of Poisson arrivals, where only the Poisson counts of the number of arrivals are independent. Additionally, the model is formulated to study the coherence of cumulative counts of the flow of Poisson readings by two independent sources. Furthermore, a model is initiated to compare the predicted value of Poisson readings by the second source based on the Poisson readings of the same Poisson flow by the first source. Numerical illustrations are provided, using simulated data and real-life data as well. The real data is on the number of car passengers recorded by traffic cameras.

CC490 Room BH (S) 2.05 MCMC METHODS AND BAYESIAN COMPUTATION Chair: Pavlo Mozharovskyi

C1538: **VariationalDCM 2.0: An updated R package for variational Bayesian inference in diagnostic classification models**
Presenter: **Kensuke Okada**, The University of Tokyo, Japan
Co-authors: Keiichiro Hijikata, Motonori Oka, Kazuhiro Yamaguchi

Variational Bayesian methods provide a computationally scalable alternative to traditional Markov chain Monte Carlo (MCMC) techniques, making them highly effective for large-scale data analyses often encountered in educational and psychological assessments. Diagnostic classification models (DCMs), a class of psychometric models, are employed to diagnose latent skills or attributes based on item response data, allowing for the assessment of specific skill mastery or non-mastery in respondents. The enhanced R package, variationalDCM, leverages these methods to efficiently and scalably estimate DCMs. It is available on CRAN, making it easily accessible for researchers. In the latest version, 2.0, several important updates have been incorporated. First, it integrates five DCMs into a unified fit function, streamlining model specification and improving usability. Second, the package now outputs results as an S3 class, enabling easy integration with other R packages and simplifying post-processing. Additionally, a newly implemented summary function offers quick access to key diagnostics and parameter estimates, enhancing the interpretability of results. The advantages of variational Bayesian inference are emphasized for DCMs, particularly its ability to deliver fast, approximate solutions with strong theoretical guarantees and demonstrate the functionality of the package through illustrative examples.

C1239: **Identification-driven MCMC**
Presenter: **Yizhou Kuang**, University of Manchester, United Kingdom
Co-authors: Toru Kitagawa

Sampling methods often face slow or non-convergence with irregular target distributions or in high-dimensional spaces. The purpose is to introduce a novel MCMC approach that leverages the knowledge of observationally equivalent sets of model parameters. It is first shown that this method performs on par with or better than conventional MCMC techniques, particularly in high-dimensional settings. It is also shown in simulation that it compares favorably to other prevalent sampling strategies, such as (adaptive) sequential Monte Carlo, especially as the dimensionality of the variables increases. For application, the performance of the algorithm is illustrated in an SVMA setting, allowing for non-invertibility.

C1447: **Augmented island resampling particle filter for particle MCMC**
Presenter: **Kari Heine**, University of Bath, United Kingdom

The ability to carry out computations in parallel is paramount to efficient implementations of computationally intensive algorithms. The applicability of the augmented island resampling particle filter (AIRPF) - an algorithm designed for parallel computing - to particle Markov chain Monte Carlo (PMCMC) is investigated. It shows that it produces a non-negative, unbiased estimator of the marginal likelihood, making it suitable for PMCMC. Moreover, the stability results previously shown for the so-called SMC algorithm to cover AIRPF are extended. As a corollary, the error of AIRPF can be bounded uniformly in time by controlling the effective number of filters, which is a diagnostic analogous to the effective sample size. Such control can be implemented by adaptively constraining the interactions between the parallel filters. The superiority of AIRPF over independent bootstrap particle filters is demonstrated not only numerically but also theoretically. In this context, the previously proposed collision analysis approach is extended to derive an explicit expression for the variance of the marginal likelihood estimate and establish an unexpected connection between the filter network topology and the marginal likelihood variance in terms of the Fibonacci sequence.

C1474: **Tuning diagonal scale matrices for HMC**
Presenter: **Jimmy Huy Tran**, University of Stavanger, Norway
Co-authors: Tore Selland Kleppe

Three approaches for adaptively tuning diagonal scale matrices for HMC are discussed and compared. The common practice of scaling according to estimated marginal standard deviations is taken as a benchmark. Scaling according to the mean log-target gradient (ISG) and a scaling method targeting the frequency of when the underlying Hamiltonian dynamics crosses the respective medians should be uniform across dimensions, which are taken as alternatives. Numerical studies suggest that the ISG method leads, in many cases, to more efficient sampling than the benchmark, particularly in cases with strong correlations or non-linear dependencies. The ISG method is also easy to implement, computationally cheap, and relatively simple to include in automatically tuned codes as an alternative to benchmark practice.

C0195: **Copula approximate Bayesian computation using distribution random forests**
Presenter: **George Karabatsos**, University of Illinois-Chicago, United States

A novel approximate Bayesian computation (ABC) framework is introduced for estimating the posterior distribution and the maximum likelihood estimate (MLE) of the parameters of models defined by intractable likelihood functions. This framework can describe the possibly skewed and high-dimensional posterior distribution by a novel multivariate copula-based distribution based on univariate marginal posterior distributions, which can account for skewness and be accurately estimated by distribution random forests (DRF) while performing automatic summary statistics (covariates) selection, and on robustly estimated copula dependence parameters. The framework employs a novel multivariate mode estimator to perform MLE and posterior mode estimation and provides an optional step to perform model selection from a given set of models with posterior probabilities estimated by DRF. The posterior distribution estimation accuracy of the ABC framework is illustrated through simulation studies involving models with analytically computable posterior distributions and exponential random graph and mechanistic network models, each defined by an intractable likelihood from which it is costly to simulate large network datasets. Also, the framework is illustrated through analyses of large real-life networks of sizes ranging between 28,000 to 65.6 million nodes (between 3 million to 1.8 billion edges), including a large multilayer network with weighted directed edges.

CC427 Room BH (SE) 2.01 TIME SERIES Chair: Eliana Christou

C0306: **Periodogram regression a two stage mixed effects approach for tropical cyclone frequency**
Presenter: **Sourav Das**, Curtin University, Australia

Co-authors: Guoqi Qian, Lyuyuan Zhang

Tropical cyclones (TC) are significant indicators of evolving climate dynamics. Two primary responses of interest are the cyclone frequency and intensity. A novel integrated modelling framework is proposed for the simultaneous modelling of TC frequency across several meteorological regions within Australasia. The key methodological insight is to model the second-order properties of multiple integer-valued time series in the frequency domain instead of parametric time domain models. A two-stage semiparametric approach is taken where large-scale environmental variation is modelled using generalized linear models while the stochastic variation, including spatial heterogeneity, is estimated using spectral analysis of time series under a hierarchical generating model. Using longitudinal data analysis, periodicities are jointly modeled in TC frequencies and their correlation with El Nino Southern Oscillation (ENSO) cycles, but also the spatial variation between regions. The fitted model is projected to obtain one-step-ahead forecasts using the principles of the best linear unbiased estimators. This semi-parametric approach avoids the uniqueness issues of parametric integer-valued time series modelling. Additional methodological advantages include tests for spatial heterogeneity and temporal second-order stationarity. The data analysis corroborates previous findings on the declining trend of tropical cyclone frequencies in the short term.

C1200: **Multi country analysis of the healthy life expectancy gap**
Presenter: **Gabriella Piscopo**, University of Naples Federico II, Italy
Co-authors: Emilia di Lorenzo, Alba Roviello, Marilena Sibillo

The progressive increase in longevity in modern society is the result of continuous progress in medicine, nutrition and technology. However, the rapid increase in life expectancy is inevitably associated with the increase in the incidence of diseases related to old age. Healthy life expectancy (HLE) is an indicator that measures the number of years individuals are expected to live free from disease or disability. Forecasting HLE is essential to ensure the stability of health systems and the sustainability of pension systems, plan the provision of health care to increasingly elderly populations, and appropriately price long-term care products. HLE is also linked to the degree of healthcare directed towards prevention and the effectiveness and efficiency of the healthcare system. In this regard, notable differences between developed and developing countries can be seen. The aim is to propose a measure of the difference between HLE in different countries, the HLE gap ratio, across some populations collected in the Global Burden of Disease database. With the aim of highlighting similarities and differences between the countries considered, a functional clustering method is applied to the multivariate time series of the HLE gap. Thus, HLE gap forecasting is carried out for some countries. The results focus on some countries with high HLE gap, suggesting the need for appropriate interventions in the healthcare system.

C1340: **Efficient outlier detection in heterogeneous time series databases**
Presenter: **Pedro Galeano**, Universidad Carlos III de Madrid, Spain
Co-authors: Daniel Pena, Ruey Tsay

A fast and powerful approach is proposed to identify univariate outliers in a large dataset of time series. The approach is highly flexible, as it can handle databases with different definitions, frequencies, and sample sizes across the marginal series. The proposed method examines the residuals of the observed series after a robust model fitting to detect outliers, uses saturated regression models to consider all observations as potential outliers, and uses the orthogonal greedy algorithm to detect significant outlying effects. The method is automatic and has been implemented to run in parallel in the texttt R package, allowing fast and efficient identification of outliers in large datasets. The performance of the proposed procedure is investigated by several simulations and the analysis of an empirical example.

C1435: **Count time series: Handling structural breaks**
Presenter: **Isabel Pereira**, University of Aveiro, Portugal
Co-authors: Magda Monteiro, Ines Pinto

Count time series are commonly used in fields like healthcare, finance, and transportation to track events such as hospital admissions or stock transactions. A well-known approach for modeling stationary count data is based on thinning operators and mimics autoregressive moving-average models. This method can generate stationary integer-valued time series with distributions like negative binomial, geometric, or Poisson. One widely studied model is the INAR(p) model, where p represents the autoregressive order. However, these models typically assume constant parameters over time, which may not always hold in practice. For instance, during an epidemic, the number of daily cases typically begins low, rises, and eventually falls. Identifying breakpoints in count processes is essential for both methodological and practical purposes, such as validating scientific hypotheses, overseeing safety-critical systems, and ensuring the accuracy of model assumptions. Considering the INAR model with structural breaks and an innovation process that accounts for overdispersion, both classical and Bayesian approaches are addressed to estimate the model parameters. The proposed methodologies are applied to a real dataset in the context of health indicators. Various techniques for detecting one or more change points are compared, utilizing both frequentist and Bayesian frameworks.

C1462: **Bayesian group-shrinkage based estimation for panel vector autoregressive models with mixed frequency data**
Presenter: **Nilanjana Chakraborty**, Indian Institute Of Management Udaipur, India
Co-authors: Kshitij Khare, George Michailidis

Panel vector auto-regressive (VAR) models are effective tools for capturing temporal relationships between a set of variables (e.g., macroeconomic indicators of an economy) while accounting for interdependencies between a set of entities (e.g., market sectors or whole economies). In the case of modelling macroeconomic data, this challenge is often further accentuated by the fact that some variables are observed at different frequencies. Existing Bayesian approaches for linking entity-specific VAR models aim to fuse relevant coefficients of the corresponding VAR transition matrices to a common value. A balanced and less stringent Bayesian approach is developed for mixed frequency panel VAR models that use group shrinkage prior distributions to borrow strength across entities but, at the same time, provide enough flexibility for entity-specific idiosyncrasies. A key novel feature is the ability to incorporate and learn the interdependence structure between various entities through an inter-entity covariance (matrix) parameter. The performance of the proposed methodology is evaluated both on synthetic data and on two economic applications; the first focuses on employment indices across neighboring US states, and the second on macroeconomic indicators of tightly integrated European economies. Finally, the theoretical properties of the modelling approach are also established.

CC482 **Room BH (SE) 2.05** **RISK ANALYSIS** **Chair: Abdelaati Daouia**

C1343: **Tail dependence in Gram-Charlier type multivariate distributions: The relevance of the moment spillovers**
Presenter: **Javier Perote**, University of Salamanca, Spain
Co-authors: Andres Mora Valencia, Ines Jimenez

The purpose is to introduce the moment interaction between different assets in the semi-nonparametric modeling of the multivariate distribution. The traditional multivariate Gram-Charlier distribution is extended by incorporating the crossed products of Hermite polynomials in the multivariate expansion. The weights of these additional terms convey spillover effects between different moments and help understand tail dependence. Bivariate portfolios are analyzed where skewness and kurtosis may interact between them and across different assets in the portfolio, showing that these new parameters may be significant pieces of information, particularly for risk measuring. Model performance for risk assessment is tested with backtesting techniques considering equally weighted portfolios of S&P 500 and Nasdaq 100 indices and major cryptocurrencies (Bitcoin and Ethereum), the latter with high-frequency data. Results show an adequate performance of the expanded Gram-Charlier multivariate distribution in terms of value-at-risk and medium shortfall, especially for high confidence levels.

C1349: **Worst-case higher moment risk measures: Addressing procyclicality and stress-testing**
Presenter: **Carlos Castro-Iragorri**, Colegio Mayor de Nuestra Senora del Rosario, Colombia

Co-authors: Fabio Gomez

The worst-case higher moment (HM) risk measure is a generalization of the expected shortfall (ES). The aim is to propose a robust solution to address distributional shifts by incorporating adaptive features into the worst-case HM risk measure. This approach offers two key advantages: first, it provides a mechanism to mitigate the inherent procyclicality of risk measures like ES; second, it allows for parametric adjustments to the risk measure, enabling the generation of diverse scenarios for stress testing. Empirical analysis using historical S&P 500 returns demonstrates that worst-case HM risk measures significantly reduce the underestimation of risk and offer more stable risk assessments throughout financial cycles compared to traditional ES predictions. These findings suggest that worst-case HM risk measures could serve as a viable alternative to regulatory add-ons for stress testing and procyclicality mitigation in financial risk management.

C1237: **Dynamic shrinkage and selection for risk factors**

Presenter: **Zheng Fan**, University of Melbourne, Australia

Financial risk factors significantly guide and shape investment decisions. Despite extensive research documenting various factors such as size and value, there is still no consensus on which factors are truly meaningful, particularly in high-dimensional dynamic modeling where their effectiveness may vary over time. The concept of the factor zoo encompasses millions of models and hundreds of distinct factors, adding complexity and substantial computational costs to the search for valuable factors. Furthermore, these models often overlook the variability in factor usefulness over time. While significant literature exists on variable selection in both cross-sectional and time-varying contexts, selections are typically assumed to be independent over time. The relevance of financial factors today is argued to be intrinsically linked to their historical trajectories, advocating for a time-dependent variable selection mechanism. Therefore, a dynamic approach is proposed that incorporates decoupling shrinkage and selection alongside Cartesian credible sets to achieve sparsity and tackle potential collinearity. This approach combines dynamic shrinkage and selection processes to effectively evaluate and identify the significance of competing risk factors over time.

C1245: **Measuring default intensities through a reduced form credit risk approach**

Presenter: **Ana Monteiro**, University of Coimbra, Portugal

Co-authors: Rui Pascoal, Mario Augusto

The probability of default (PD) is a critical metric in credit risk analysis, representing the likelihood that a borrower will fail to fulfil their debt obligations. Estimating PD can be approached through structural or reduced-form models. The focus is on the estimation of default probabilities under the risk-neutral measure using the Jarrow/Turnbull model, a reduced-form approach, by applying it to both simulated and market data for some companies such as Apple and Beyond Meat. The objective is to extract valuable insights into market dynamics and the impact of historical events on credit risk. To derive relevant insights into bond dynamics, the analysis incorporates zero-coupon yields obtained through the Svensson approach, a method commonly employed by central banks to infer default-free zero-coupon bond prices. In addition, coupon maturities, cash flows and market bond prices are used for the cited companies, providing a comprehensive view of their respective credit spread term structures. The efficiency of the Jarrow/Turnbull models estimation method is first evaluated by comparing the discrepancies between estimated and simulated bond prices. Subsequently, the analysis extends to a comparative evaluation of the estimated default probabilities for the cited companies. The findings aim to enhance investors' understanding of the factors driving default risk, thereby enabling more informed decision-making in the bond market.

Sunday 15.12.2024 08:45 - 10:25 Parallel Session F – CFECMStatistics2024

CI052 Room Auditorium RECENT ADVANCES IN CLUSTERING Chair: Matthieu Marbac

C0187: **Choosing the number of biological species in the presence of spatial patterns of differentiation**
Presenter: **Gabriele d Angella**, University of Bologna, Italy
Co-authors: Christian Hennig

The delimitation of biological species and the identification of their diverse subpopulations are key activities for the preservation of biodiversity. Statistical delimitation methods use empirical data to suggest how many species are represented in a dataset. This task is hard in that spatial patterns of differentiation can introduce genetic variation between populations belonging to the same species. Software packages that consider geographic information to estimate ancestry proportions from spatially explicit genotypic data can be used to delimit species in these setups. However, determining the number of species represented in a dataset remains a challenge, and practitioners are often left with heuristic methods that are not mathematically well-defined. Therefore, it can be beneficial to develop techniques that inform this choice. Options include methods that study the relationship between genetic and geographic dissimilarities between individuals, routines that contrast the delimitation software output with appropriate null models and approaches from selective inference. The effectiveness of these methodologies can be assessed on data generated with SLiM, a software package that can simulate spatially explicit genetic data.

C0538: **Biclustering listeners and music genres using a composite likelihood-based approach**
Presenter: **Francesca Martella**, La Sapienza University of Rome, Italy
Co-authors: Monia Ranalli

A finite mixture model is proposed that simultaneously clusters listeners' habits and music genres. By following the underlying response variable (URV) approach, the music genres are treated as discretized versions of latent continuous variables, which are distributed according to a mixture of Gaussians. To introduce a partition of the music genres within each mixture component, a factorial representation of the data is used, where a binary row stochastic matrix represents music genre membership. This method allows associating each mixture component with a cluster of music genres, thereby defining a bicluster of listeners' habits and music genres. Given the numerical complexity of the likelihood function, model parameters are estimated using a composite likelihood (CL) approach, leading to a computationally efficient, like EM algorithm. The results illustrate the effectiveness of the proposed model in discovering significant patterns within the data. The model adeptly identifies clusters of listeners who share similar preferences for clusters of music genres, revealing both the listener groups with common tastes and the relationships between different music genres within these groups. Additionally, by allowing the number of clusters of music genres to vary with listener clusters, the model adeptly captures the inherent variability in listener preferences, exhibiting its flexibility and accuracy in representing the data and uncovering interesting patterns in listener behavior.

C0217: **Mixture models via continuous sparse regression**
Presenter: **Yohann De Castro**, Institut Camille Jordan, France

Mixture models are a popular family of model-based clustering methods. One prominent example is given by the expectation-maximization (EM) algorithm in Gaussian mixture models. Taking advantage of recent advances in continuous sparse regression, a new method is introduced, referred to as Beurling LASSO, and it is shown that one can recover (i) The number of components, (ii) The locations of the mixture at a rate of $n^{-1/4}$, (iii) The weights of the mixture at a parametric rate of $n^{-1/2}$, where n is the sample size. When the sample size is large, it is proven that one can reduce the dimensionality of data while preserving important information for clustering. This compressed representation, called a sketch, is significantly smaller than the original data but still retains enough information for our method to operate effectively. It is proven that the sketch size does not depend on the sample size but rather on the number of components and the dimension.

CO381 Room S-2.25 RECENT ADVANCES IN THEORY AND METHODS OF DEPENDENT DATA Chair: Wei-Ying Wu

C0728: **Pseudo-spectra of multivariate inhomogeneous spatial point processes**
Presenter: **Junho Yang**, Academia Sinica, Taiwan

Spectral analysis is a technique that characterizes the second-order structure of stationary time series, random fields, and point processes. However, in spatial point processes, a stationary assumption, especially homogeneity of the first-order intensity assumption, is often considered too stringent. A new spectral analysis is proposed for a multivariate inhomogeneous point process observed on $\mathbb{R}^d$. A key idea is the asymptotic behavior of the discrete Fourier transform (DFT) of the observed spatial point pattern and its tapered version. It is shown that, even in the case of inhomogeneous processes, the expectation of the periodogram converges to some deterministic matrix-valued function, which is Hermitian and positive definite. This limit is referred to as the pseudo-spectrum. It is shown that the defined pseudo-spectrum has an interpretation in terms of integrating the local spectra. The consistent estimator of the pseudo-spectrum is derived via periodogram smoothing, and methods are proposed to select the bandwidth. Finally, the proposed estimator is demonstrated through simulations and real data analysis.

C0850: **Multi-resolution spatial methods on the sphere**
Presenter: **HaoYun Huang**, National Dong Hwa University, Taiwan

Spatial prediction is investigated on the sphere in the presence of measurement errors, where data may be irregularly located. A special class of basis functions are first developed in the thin-plate-spline (TPS) function space on the sphere. These basis functions are ordered according to their level of smoothness from large-scale features to small-scale details, providing a multi-resolution representation and an orthogonal transformation of the data. Theoretically, it is shown that the number of basis functions selected by the conditional Akaike information criterion is small, and the resulting reduced-rank estimate achieves a good convergence rate to the target function. In addition, a multi-resolution mixed-effects spatial model is developed on the sphere by including a Gaussian spatial process to capture fine-scale information. Since large-scale features are captured by leading basis functions, the small-scale spatial process tends to have a short spatial dependence range, leading to a universal kriging estimate that allows rapid computations. A simulation experiment is performed, and an application to global sea-surface-temperature data observed from a satellite is given to demonstrate the effectiveness of the proposed method.

C1155: **Modeling longitudinal area data with zero-modified via GEE**
Presenter: **Hong-Ding Yang**, National Chiayi University, Taiwan

Benefiting from the assumptions of the hurdle model, a parameter estimation method is proposed under the generation mechanism of repeated measurements of area data with zero-modified. However, the spatiotemporal correlation of data is unknown in practice; generalized estimating equations (GEE) are used to estimate the regression coefficients. GEE exhibits robustness even when the underlying distribution of the reaction is unknown. Therefore, it is assumed that the marginal distribution of responses follows a hurdle binomial distribution, accommodating the zero-inflation and zero-contraction cases. Furthermore, an iterative non-parametric technique is employed to update the working correlation matrix and utilize a jackknife approach to approximate the estimated variance of GEE, resulting in more valid and reliable interval estimates. Numerical results show that the proposed method is promising for analyzing complex spatiotemporal data with zero-modification characteristics. Comparative analyses demonstrate the superiority of the approach over alternative methods based on mean squared error measurements.

C1209: **Tail estimation of the spectral density under fixed domain asymptotics**
Presenter: **Wei-Ying Wu**, National Dong Hwa University, Taiwan

Many estimation studies have been conducted using fixed-domain asymptotics for spatial processes. However, most have been studied under the spatial domain approach, in which specific covariance models are assumed. Unlike the spatial domain approach, spectral density is one way to describe the spatial dependence for weakly stationary spatial processes. A methodology is proposed to simultaneously estimate parameters that describe the tail behavior of spectral densities under fixed domain asymptotics. The spectral tail parameters are defined by assuming only a tail structure of the spectral densities, which can involve a broader class of spatial dependence models. Theoretical properties of the proposed estimator, such as consistency and asymptotic results, are introduced. Meanwhile, simulation experiments with real data studies have also been shown to support theoretical studies.

CO103 Room S-1.01 HITEC: RECENT PROGRESS IN HIGH-DIMENSIONAL TIME SERIES Chair: Marc Hallin

C0300: **Optimal tests for the absence of random individual effects in large n and small T dynamic panels**
Presenter: **Yuichi Goto**, Kyushu University, Japan
Co-authors: Nezar Bennala, Marc Hallin

The problem of testing the absence of individual effects is considered in dynamic panels with AR(p) disturbances. A local asymptotic normality property is established for the number of individuals tending to infinity and the time series length fixed. Then, locally asymptotically optimal parametric and pseudo-Gaussian tests are constructed.

C0366: **Weak factors are everywhere**
Presenter: **Philipp Gersing**, University of Vienna, Austria
Co-authors: Christoph Rust, Manfred Deistler, Matteo Barigozzi

There are two different approaches to time series factor models: a) The approximate factor model, where the factors are loaded contemporaneously to the common component. b) The generalised dynamic factor model (GDFM), where the factors are loaded with lags. By introducing the canonical decomposition of factor models, it is shown how both approaches are related, and their conceptual difference is clarified: Both models entail two different types of common/idiosyncratic components, respectively. The canonical decomposition includes what is called the weak common component, which is the difference between the dynamic- and the static common component. It is driven by potentially infinitely many non-pervasive factors, i.e., weak factors (not to be confused with factors being pervasive but at a slower rate). It exemplifies why these types of weak factors should not be neglected in theory and practice. Furthermore, a simple estimator for the canonical decomposition is proposed and applied to US macroeconomic data. The estimates reveal that the weak common component can account for up to 20% of the total variation of individual variables.

C0439: **Global bank network connectedness revisited: What is common, idiosyncratic and when**
Presenter: **Luca Margaritella**, Lund University, Sweden
Co-authors: Jonas Krampe

The problem of estimating high-dimensional global bank network connectedness, both in the time and frequency domain, is revisited. Instead of directly regularizing the high-dimensional vector of realized volatilities as in Demirer et al. (2018), we estimate a dynamic factor model with sparse VAR idiosyncratic components. This allows to disentangle: (I) the part of system-wide connectedness (SWC) due to the common component shocks (the banking market), and (II) the part due to the idiosyncratic shocks(the single banks). Via spectral density estimation, SWC, (I) and (II) can be further decomposed into short (S), medium (M), long (L) frequency responses to shocks. We employ both the original dataset as in Demirer et al. (2018) (daily data, 2003-2013), as well as a more recent vintage (2014-2023). For both, we compute SWC due to (I), (II), (I+II) and (S=monthly), (M=quarterly), (L=yearly), providing bootstrap confidence bands. We find SWC to spike upward during global crises, disentangling how this is primarily driven by (I), (S). In normal times instead, SWC is primarily driven by (II), (M).

C0516: **Spectral-based variable selection of high-dimensional data for prediction of the El Nino/Southern Oscillation cycle**
Presenter: **Alessandro Giovannelli**, University of L'Aquila, Italy
Co-authors: Tommaso Proietti

The El Nino/Southern Oscillation (ENSO) phenomenon is a key driver of interannual climate variability. A novel procedure is introduced based on large dynamic factor models (DFMs) to improve the prediction of sea surface temperatures in the four El Nino regions. A significant body of literature on DFMs addresses the selection of variables for factor estimation, which directly impacts predictive accuracy. Existing methodologies for constructing targeted principal components often rely on static correlation for variable selection. The approach departs from these methodologies by proposing a new selection procedure that screens variables based on significant correlation within the frequency band most relevant to the variable of interest. This method, known as dynamic correlation, assesses co-movements between the target variable and covariates within the frequency band of interest, capturing dynamic correlations over time. This methodology is applied to a high-dimensional dataset containing sea surface temperatures from the Nino regions, using El Nino 3.4 as the target variable. To evaluate the effectiveness of the procedure, a real-time exercise is conducted comparing results obtained using the targeted dataset against those obtained using all available series. Preliminary results indicate significant potential to improve ENSO prediction accuracy.

CO058 Room S-1.04 RECENT DEVELOPMENTS IN STATISTICAL MODELING FOR STOCHASTIC PROCESSES Chair: Masayuki Uchida

C0370: **Statistical inference for generalized Gerber-Shiu functions in risk theory**
Presenter: **Yasutaka Shimizu**, Waseda University, Japan

The Gerber-Shiu function is an important ruin-related quantity in risk theory, which does not generally have an explicit form. If the underlying asset process is a spectrally negative Levy process, it has an integral representation via the scale function of the Levy process. A novel series representation is introduced for this scale function, employing Laguerre polynomials to construct a uniformly convergent approximate sequence. Additionally, statistical inference of the Gerber-Shiu functions is derived based on specific discrete observations and presenting estimators.

C0446: **Parametric estimation for a linear parabolic SPDE in two space dimensions under small diffusivity asymptotics**
Presenter: **Yozo Tonaki**, Osaka University, Japan
Co-authors: Yusuke Kaino, Masayuki Uchida

Parametric estimation is considered for a second-order linear parabolic stochastic partial differential equation (SPDE) in two space dimensions driven by a Q-Wiener process under small diffusivity asymptotics. An estimator of the reaction parameter is first provided in the linear parabolic SPDE in two space dimensions with small diffusive and advective parameters based on continuous spatiotemporal data applying the methodology of an existing study to the SPDE in two space dimensions. An estimator of the reaction parameter is then constructed based on high-frequency spatiotemporal data by discretizing the estimator based on the continuous data. Furthermore, it is shown that the estimators have consistency and asymptotic normality under certain asymptotic conditions, and the asymptotic properties of the estimator are verified based on high-frequency data by numerical simulations.

C1054: **Statistical analysis of a stochastic boundary model for high-frequency data from a limit order book**
Presenter: **Markus Bibinger**, University of Wurzburg, Germany

Statistical methods to infer characteristics of a semimartingale efficient log-price process are presented. Observations are from a boundary model with one-sided microstructure noise for high-frequency prices of limit orders. The previous one-dimensional model is generalized to a multivariate model, and the focus is on the estimation of the covolatility matrix. Asymptotic results are established in a high-frequency regime. Convergence rates are shown to be faster than under standard market microstructure noise. They hinge on a tail index of the noise distribution. The estimation of this noise tail index and adaptive inference on the semimartingale is addressed.

C0413: **Log-rank test with coarsened exact matching**
Presenter: **Nakahiro Yoshida**, University of Tokyo, Japan

It is of special importance in the clinical trial to compare survival times between the treatment group and the control group. In practice, the distributions of the covariates differ between the two groups since the distribution of the assignments depends on the covariates of the individuals. The complex structure between covariates, treatment assignment and survival time makes it difficult to apply parametric methods, such as the propensity score by logistic regression, to correctly match individuals of the different groups to assess the treatment effect. The coarsened exact matching (CEM) is considered, which does not need any parametric models, and a weighted log-rank statistic based on CEM is proposed. Asymptotic properties of the weighted log-rank statistic are given, i.e., the asymptotic normality under the null hypothesis and the consistency of the test. Simulation studies show that the log-rank statistic with CEM is more robust than the log-rank statistic with the propensity score matching.

CO097 **Room S-1.06** CONTEMPORARY DIRECTIONAL STATISTICS **Chair: Andriette Bekker**

C0376: **Bayesian inference for sphere-on-sphere regression with optimal transport map**
Presenter: **Tin Lok James Ng**, Trinity College Dublin, Ireland
Co-authors: Andrew Zammit Mangion, Kwok-Kun Kwong, Jiakun Liu

The field of spherical regression, where both covariate and response variables take values on the sphere, has seen extensive methodological development over time. Despite the creation of various parametric and non-parametric techniques to tackle spherical regression, it remains a challenging problem due to the complexities involved in parameterizing regression models between spherical domains. Additionally, there is a notable gap in methods for quantifying uncertainties associated with the estimated regression maps. To address these challenges, optimal transport theory is utilized, and a Bayesian approach is employed. This framework eliminates the necessity for directly parameterizing the regression map and enables uncertainty quantification.

C0545: **A data-driven smoothing parameter for circular kernel density estimation**
Presenter: **Jose Ameijeiras-Alonso**, Universidade de Santiago de Compostela, Spain

A novel data-driven smoothing parameter is introduced specifically designed for circular kernel density estimation. By adapting the well-established Sheather and Jones bandwidths to the circular domain, unknown parameters are replaced with estimates derived through plug-in techniques. Theoretical support is provided for the method, deriving the asymptotic mean squared error of the density estimator, its derivatives, and its functionals for circular data. A comprehensive simulation study demonstrates the superior performance of the proposed selectors compared to existing data-driven smoothing parameters. Additionally, the practical application of the plug-in rules is illustrated using real-world data on the timing of car accidents.

C1214: **On regime changes in text data using hidden Markov model of contaminated von Mises-Fisher distribution**
Presenter: **Shuchismita Sarkar**, Bowling Green State University, United States
Co-authors: Yingying Zhang, Yuanyuan Chen, Xuwen Zhu

The purpose is to present a novel methodology for analyzing temporal directional data with scatter and heavy tails. A hidden Markov model with contaminated von Mises-Fisher emission distribution is developed. The model is implemented using a forward and backward selection approach that provides additional flexibility for contaminated as well as non-contaminated data. The utility of the method for finding homogeneous time blocks (regimes) is demonstrated in several experimental settings and two real-life text data sets containing presidential addresses and corporate financial statements, respectively.

C1513: **Insights into the construction of an alternative bivariate cardioid distribution**
Presenter: **JT Ferreira**, University of Pretoria, South Africa
Co-authors: Andriette Bekker, Delene van Wyk de Ridder

Bivariate generalizations of the well-known cardioid distribution have been investigated in the literature by relying on a mixture approach. Specifically, marginal (univariate) cardioid distributions are blended together based on the concentration parameter assumed to be beta-distributed. Alternative choices are investigated for the distribution of the concentration parameter and the resultant effect on the probabilistic behavior of the circle emanating from the corresponding developed alternative bivariate cardioid distributions. Some key theoretical aspects are considered and are accompanied by numerical illustrations and results.

CO314 **Room S-1.27** RECENT ADVANCES OF STATISTICAL INFERENCE **Chair: Yang Han**

C0852: **Prediction and calibration for all future values: Simultaneous tolerance regions for multivariate regression**
Presenter: **Lingjiao Wang**, University of Manchester, United Kingdom
Co-authors: Yang Han, Wei Liu, Frank Bretz

Multiple-use prediction and calibration for all future values play a valuable role in many fields, including industry, sports, health and medical research. Simultaneous tolerance regions (STRs) can be used for this purpose in multivariate regression, but no construction methods are currently available in the literature. The first aim is to fill the gap by constructing STRs and providing a solution to multiple-use prediction and calibration problems. We also develop weighted simultaneous tolerance regions (WSTRs), showing that confidence sets based on WSTRs can exactly satisfy the key property for multiple-use calibration. Through the lens of WSTRs, we address the misconception that the confidence sets derived from pointwise tolerance regions can guarantee the key property. Several methods are proposed for computing the critical constants for STRs and WSTRs. Simulation studies are conducted to compare these methods in terms of accuracy and computational cost. Real-data examples are provided for illustration.

C0896: **Random interval distillation for detection of change-points in Markov chain Bernoulli networks**
Presenter: **Xinyuan Fan**, Tsinghua University, China
Co-authors: Weichi Wu

A new and generic approach for detecting multiple change-points in dynamic networks with Markov formation, termed random interval distillation (RID). By collecting random intervals with sufficient strength of signals and reassembling them into a sequence of informative short intervals, together with universal singular value thresholding, the new approach can achieve a nearly minimax optimality as their independent counterparts for both detection and localization bounds in low-rank networks without any prior knowledge about minimal spacing, which is unlike many previous methods. In particular, motivated by a recent nonasymptotic bound, the method utilizes the operator norm of CUSUMs of the adjacency matrices, achieving the aforementioned optimality without sample splitting as required by the previous method. For practical applications, a clustering-based and data-driven procedure is introduced to determine the optimal threshold for signal strength, utilizing the connection between RID and clustering. The effectiveness and usefulness of the methodology are examined via extensive simulation studies and a real data example.

C0936: **Uniform variance reduced simultaneous inference of time-varying correlation networks**
Presenter: **Lujia Bai**, Tsinghua University, China
Co-authors: Weichi Wu

A flexible framework is proposed for inferring large-scale time-varying and time-lagged correlation networks from non-stationary multivariate or high-dimensional non-stationary time series with piecewise smooth trends. Built on a novel and unified multiple-testing procedure of time-lagged cross-correlation functions with a fixed or diverging number of lags, the method can accurately disclose flexible time-varying network structures associated with complex functional structures at all time points. The applicability of the method is broadened to the structure breaks by developing difference-based nonparametric estimators of cross-correlations, accurate family-wise error control is achieved via a bootstrap-assisted procedure adaptive to the complex temporal dynamics, and the probability of recovering the time-varying network structures is enhanced using a new uniform variance reduction technique for simultaneous inference of nonparametric estimate, which is of independent interest. The asymptotic validity of the proposed method is proved, and its effectiveness is demonstrated in finite samples through simulation studies and empirical applications.

C1304: **Fully functional sieve covariance inference of locally stationary functional time series**
Presenter: **Yan Cui**, Reed College, United States

Simultaneous statistical inference is established for the auto-covariance functions of locally stationary functional time series based on full functional information rather than employing dimension reduction techniques. The sieve method is leveraged to estimate the unknown auto-covariance function with flexible choices of orthonormal basis functions. A fully functional multiplier bootstrap methodology is proposed to construct asymptotically correct simultaneous confidence regions for the auto-covariance functions, which can be validated by a uniform Gaussian approximation over all Euclidean convex sets for sums of a class of moderately high-dimensional locally stationary time series. The proposed approach is applied to an air quality functional time series dataset to investigate the variability of the auto-covariance functions.

CO095 Room K0.16 NEW DEVELOPMENTS IN STATISTICAL NETWORK ANALYSIS Chair: Jonathan Stewart

C0570: **Detecting planted partition in sparse multi-layer networks**
Presenter: **Sagnik Nandy**, University of Chicago, United States
Co-authors: Anirban Chatterjee, Ritwik Sadhu

Multi-layer networks represent the interdependence between the relational data of individuals interacting with each other via different types of relationships. To study the information-theoretic phase transitions in detecting the presence of planted partition among the nodes of a multi-layer network with additional node covariate information and diverging average degree, a recent study introduced a multi-layer contextual stochastic block model. The problem of detecting planted partitions is considered in the multi-layer contextual stochastic block model when the average node degrees for each network are greater than 1. The sharp phase transition threshold is established for detecting such planted bi-partition. Above the phase-transition threshold, testing the presence of a bi-partition is possible, whereas, below the threshold, no procedure to identify the planted bi-partition can perform better than random guessing. The derived detection threshold is further established to coincide with the threshold for weak recovery of the partition, and a quasi-polynomial time algorithm is provided to estimate it.

C0708: **A spike-and-slab prior for dimension selection in generalized linear network eigenmodels**
Presenter: **Joshua Loyal**, Florida State University, United States
Co-authors: Yuguo Chen

Latent space models (LSMs) are frequently used to model network data by embedding a network's nodes into a low-dimensional latent space; however, choosing the dimension of this space remains a challenge. To this end, the aim is to begin by formalizing a class of LSMs called generalized linear network eigenmodels (GLNEMs) that can model various edge types (binary, ordinal, non-negative continuous) found in scientific applications. This model class subsumes the traditional eigenmodel by embedding it in a generalized linear model with an exponential dispersion family random component and fixes identifiability issues that hindered interpretability. Next, a Bayesian approach is proposed to dimension selection for GLNEMs based on an ordered spike-and-slab prior that provides improved dimension estimation and satisfies several appealing theoretical properties. It is shown that the model's posterior is consistent and concentrates on low-dimensional models near the truth. The approach's consistent dimension selection is demonstrated on simulated networks. Lastly, GLNEMs are used to study the effect of covariates on the formation of networks from biology, ecology, and economics and the existence of residual latent structure.

C0764: **Learning cross-layer dependence structure in multilayer networks**
Presenter: **Jiaheng Li**, Florida State University, United States
Co-authors: Jonathan Stewart

Multilayer networks are a network data structure in which elements in a population of interest have multiple modes of interaction or relation, represented by multiple networks called layers. A novel class of models is proposed for cross-layer dependence in multilayer networks, aiming to learn how interactions in one or more layers may influence interactions in other layers of the multilayer network by developing a class of network separable models which separate the network formation process from the layer formation process. In the framework, existing single-layer network models are extended to a multilayer network model with cross-layer dependence. Non-asymptotic bounds are established on the error of estimators and demonstrate rates of convergence for both maximum likelihood estimators and maximum pseudolikelihood estimators in scenarios of increasing parameter dimension. Non-asymptotic error bounds are additionally established on the multivariate normal approximation and elaborate a method for model selection which controls the false discovery rate. Simulation studies are conducted which demonstrate that the framework and method work well in realistic settings which might be encountered in applications. Lastly, the utility of the method is illustrated through an application to the Lazega lawyers network.

C1125: **A regression framework for studying relationships among attributes under network interference: Statistical theory**
Presenter: **Michael Schweinberger**, The Pennsylvania State University, United States
Co-authors: Cornelius Fritz, Subhankar Bhadra, David Hunter

When network data are collected, the structure of networks is often of secondary interest compared to the question of how networks affect individual or collective outcomes. The well-established class of models known as generalized linear models (GLMs), which includes linear and logistic regression, assumes that the response of a given unit depends on predictors measured on that unit but is unaffected by predictors and responses of other units. A statistical framework that captures complex and realistic dependencies among attributes and connections is introduced while retaining the virtues of GLMs. The framework helps study relationships among attributes under network interference and is applicable to binary, count-valued, and real-valued attributes. Theoretical guarantees are established based on a single observation of dependent attributes and connections. In a companion talk, it is demonstrated that the framework is amenable to scalable statistical computing.

CO222 Room K0.18 STATISTICAL METHODS FOR ENVIRONMENTAL SCIENCES Chair: Julia Schaumburg

C0204: **Trends in temperature data: Micro-foundations of their nature**
Presenter: **Jesus Gonzalo**, Universidad Carlos III de Madrid, Spain

Co-authors: Lola Gadea, Andrey Ramos

Determining whether global average temperature (GAT) is an integrated process of order 1, I(1), or a stationary process around a trend function is crucial for detection, attribution, impact, and forecasting studies of climate change. The nature of trends is investigated in GAT building on the analysis of individual temperature grids. The micro-founded evidence suggests that GAT is stationary around a non-linear deterministic trend in the form of a linear function with one structural break. This break can be attributed to a combination of breaks on individual grids and the standard aggregation method under acceleration in global warming

C0303: **Spatiotemporal modeling for record-breaking temperature events**
Presenter: **Ana C Cebrian**, University of Zaragoza, Spain
Co-authors: Jorge Castillo-Mateo, Alan Gelfand, Jesus Asin, Zeus Gracia

The occurrence of record-breaking temperature events is one of the evidence of climate change. In this context, an approach is presented to investigate the occurrence of record-breaking temperatures across years for any given day within the year within a space-time framework. Formal statistical analysis of record-breaking events has primarily been developed within the probability community, using results from the stationary record-breaking setting. However, that framework is not sufficient for analyzing actual record-breaking data, which requires rich modeling of the indicator events defining record-breaking series. A dataset consisting of over sixty years of daily maximum temperatures across Peninsular Spain is used. A novel and thorough exploratory data analysis leads to the proposal of hierarchical conditional models for the indicator events. The final model includes an explicit trend, necessary autoregression terms, spatial behavior captured by the distance to the coast, useful interactions, helpful spatial random effects, and very strong daily random effects. The fitted model shows that global warming trends have increased the number of records expected in the past decade almost two-fold, but it also estimates highly differentiated climate warming rates in space and by season.

C0705: **Nowcasting of high precipitation events with deep learning**
Presenter: **Yuliya Shapovalova**, Radboud University, Netherlands

Precipitation nowcasting is critical for weather-dependent decisions but remains challenging despite active research. Combining radar data and deep learning has created new research opportunities. Radar data, with high space-time resolution, are ideal for nowcasting, while deep learning exploits possible nonlinearities in the precipitation process. Deep learning approaches have matched or outperformed optical flow methods for low-intensity precipitation, but high-intensity events nowcasting remains difficult. Two different deep learning architectures, deep generative model and diffusion model, are built upon and are extended to enhance the nowcasting of heavy precipitation. Specifically, different loss functions and the effect of adding temperature data as an additional feature are explored. Using KNMI radar data and 590-min lead times, the model with these enhancements outperforms state-of-the-art models, effectively nowcasting high rainfall intensities up to 60-min lead times.

C0791: **Modelling temperature persistence using seasonal quantile autoregressions**
Presenter: **Barend Spanjers**, Vrije Universitent and Tinbergen Institute, Netherlands

A quantile autoregressive model is proposed that incorporates trends and seasonal components in both the quantile-specific constants and autoregressive coefficients to model temperature persistence. By allowing the quantile constant and persistence coefficient to vary over time, it is found that the persistence coefficients for the warmer quantiles are higher in summer (compared to the colder quantiles), while the lower quantiles show higher persistence in winter (compared to the warmer quantiles). This persistency ranking of the quantile-specific coefficients reverses in spring and again in autumn. Results show that the autoregressive coefficients of the median have not significantly changed, whereas the coefficients have significantly increased for the high quantiles and decreased for the low quantiles in western and northern Europe. These results are compared with changes in atmospheric circulation.

CO323 Room K0.19 BAYESIAN METHODS FOR EXTREME VALUES Chair: Ramses Mena

C1140: **Construction, estimation and application of diffusion processes for extreme values**
Presenter: **Consuelo Nava**, University of Aosta Valley, Italy
Co-authors: Ramses Mena

A simple yet powerful method is proposed to construct strictly stationary Markovian models with given but arbitrary invariant distributions. The idea is based on a Poisson transform modulating the dependence structure in the model. An appealing feature of the approach is the ability to fully control the underlying transition probabilities and, therefore, incorporate them within standard estimation methods. Given the proposed representation of the transition density, a Gibbs sampler algorithm based on the slice method is proposed and implemented. The construction results from a Bayesian perspective. In the discrete-time case, special attention is placed on the class of generalized inverse Gaussian (GIG) distributions. The GIG class is very flexible and allows one to obtain various explicit results: it is an interesting choice, especially for econometric or financial applications involving extreme values.

C1224: **Tuning-free objective Bayesian inference for extremes**
Presenter: **Paolo Onorati**, University of Padova, Italy
Co-authors: Isadora Antoniano-Villalobos, Antonio Canale

The generalized extreme value (GEV) distribution is widely used for modeling extreme events. Despite its frequent application, there is no consensus on the choice of prior distributions for the parameters in the Bayesian framework. The aim is to propose the usage of an objective prior based on a scoring rule, leading to a multivariate Lomax distribution for positive parameters and a double multivariate Lomax distribution for the general case. However, despite being well motivated by theoretical arguments, this choice introduces computational challenges, as the full conditional distributions do not have a closed form. Practical solutions are provided to these issues by exploiting a generalized elliptical slice sampling (GESS), which yields a self-contained algorithm that does not require tuning parameters and is rejection-free, ensuring that consecutive values in the chain are distinct. This method is quite general and can be extended to many other parametric families. The approach utilizes a new version of the multivariate logistic distribution, represented as a scale mixture of Gaussian distributions. It is shown how to compute its density and sample from the full conditional of the mixing density. The algorithm's performance is evaluated through simulation studies and empirical analysis.

C1170: **Enhancing intensity-duration-frequency curves estimation: A Markov dependence approach**
Presenter: **Isadora Antoniano-Villalobos**, Ca Foscari University of Venice, Italy
Co-authors: Mehwish Zaman, Ilaria Prosdocimi

A fundamental challenge in hydrological risk assessment is the accurate estimation of the probabilities of rainfall events exceeding given (typically high) intensities. Since the design of infrastructures must take into account rainfall extremes at different temporal scales, coherence between estimated exceedance probabilities across different durations is desirable. Intensity duration frequency (IDF) curves, which describe the expected frequency of extreme rainfall intensities measured at different durations, are an important tool in this context. Most existing methods for IDF estimation introduce adequate shape constraints through duration-dependent generalized extreme value (dGEV) distributions but assume that rainfall accumulations are independent across durations. While this assumption does not seem realistic, the introduction of dependence in existing models comes at a high computational cost and often requires approximate inference. An alternative model is proposed based on a first-order Markov assumption, which incorporates dependence between consecutive (equally spaced) rainfall durations via bivariate generalized extreme value (GEV) distributions, while the marginal distributions are dGEV, satisfying shape constraints. The flexibility of the model depends on the family of bivariate distributions controlling the dependence, which can be parametric or nonparametric.

C1362: **A Bayesian Lasso for tail index regression**
Presenter: **Miguel de Carvalho**, University of Edinburgh and Universidade de Aveiro, Portugal
Co-authors: Daniel Paulin

Extreme events, such as heatwaves, wildfires, and widespread flooding, can have devastating impacts on communities and ecosystems. The purpose is to introduce a novel regression model for heavy-tailed phenomena, leveraging Bayesian regularization within a generalized additive framework. This approach is centered on a conditional Pareto-type specification, enhanced by Bayesian Lasso shrinkage priors and refined with low-rank thin plate splines basis expansion. The effectiveness of the proposed methods is demonstrated using both artificial and real data, with a focus on extreme wildfire events in Portugal and the key factors driving them.

CO294 Room K0.20 DISTANCE AND DEPTH BASED METHODS FOR DATA SCIENCE Chair: Silvia Salini

C1145: **Robust fast k-medoids for large mixed-type data**
Presenter: **Aurea Grane Chavez**, Universidad Carlos III de Madrid, Spain
Co-authors: Fabio Scielzo

New robust clustering algorithms for large datasets of mixed-type data are proposed. Their performance is analyzed through an extensive simulation study and compared to a wide range of existing clustering alternatives in terms of both predictive power and computational efficiency. MDS is used to visualize clustering results.

C1152: **Robust distance-based generalized linear models: Some proposals**
Presenter: **Eva Boj**, Universidad Carlos III de Madrid, Spain
Co-authors: Aurea Grane Chavez, Agustin Mayo-Iscar

Distance-based generalized linear models are prediction tools that can be applied to any kind of data whenever a distance measure is computed among units. Robust ad-hoc metrics are proposed to be used in the predictors' space of these models, incorporating more flexibility into this tool. Their performance is evaluated by means of a simulation study and several applications of real data are provided. Computations are made using the dbstats package for R.

C1380: **Converting textual data into structured survey data: A ChatGPT approach**
Presenter: **Giancarlo Manzi**, University of Milan, Italy
Co-authors: Aurea Grane Chavez, Qi Guo

A new approach to transforming textual data into survey data with the use of chatbot technologies is presented. ChatGPT APIs are considered to associate each respondent's response to an open-ended question (included in a 2019 survey questionnaire about a bike-sharing service in Milan, Italy) to the most probable Likert scale question in the questionnaire. ChatGPT is also asked to give, according to the meaning of each of such responses, a possible rating answer on a Likert scale for the chosen question. In this way, the congruence of the answers is evaluated to the open-ended question to the Likert scale options chosen by respondents and form a reliability measure to be compared to standard questionnaire reliability measure s like test-retest reliability, inter-rate reliability, etc.

C1131: **Computation of non-zero empirical expectile depths**
Presenter: **Maicol Ochoa**, Universidad Carlos III de Madrid, Spain
Co-authors: Ignacio Cascos, Ha Thi Khanh Linh

In nonparametric multivariate statistics, the depth of a point indicates its degree of centrality relative to a data cloud. A higher depth means greater centrality. Several depth concepts, such as halfspace, simplicial, and zonoid depths, have gained popularity due to their utility in outlier detection, supervised and unsupervised classification, and process monitoring. However, the empirical versions of such depth notions assume a value of zero for any point outside the convex hull of the data cloud, which poses challenges for real-world applications. To address this limitation, the focus is on the expectile depth and introduction of related empirical constructions that remain strictly positive for any point, regardless of its position. Discussion includes the computation of these constructions in any dimension.

CO087 Room K0.50 DESIGN OF EXPERIMENTS FOR COMPLEX DATA Chair: Kalliopi Mylona

C0272: **Variable shrinkage and subdata selection in big data**
Presenter: **Vasilis Chasiotis**, Athens University of Economics and Business, Greece
Co-authors: Lin Wang, Dimitris Karlis

In the field of big data analytics, the search for efficient subdata selection methods that enable robust statistical inferences with minimal computational resources is of high importance. A procedure prior to subdata selection could perform variable selection, as only a subset of a large number of variables is active. An approach is proposed for situations where both the size of the full dataset and the number of variables are large. This approach identifies the active variables by applying a procedure inspired by the random Lasso, followed by subdata selection based on leverage scores, in order to build a predictive model. The proposed approach outperforms existing methods in the current literature in both variable selection and prediction while also exhibiting significant improvements in computing time. The usage of the full dataset is considered as well. Simulation experiments, as well as a real data application, are provided.

C0278: **I-optimal robust Bayesian designs to control model misspecification**
Presenter: **Irene Garcia-C. Gutierrez**, University of Castilla-La Mancha, Spain
Co-authors: Kalliopi Mylona

Robust design techniques are crucial in experiments where the behavior of the response is poorly known prior to running the experiments. Although there are numerous alternatives to address this problem, the focus is on the approach introduced by an existing study, which used mean square error as a measure of design quality to control the bias introduced by model misspecification. This approach is quite general in the sense of covering a wide range of possible responses. Nevertheless, a strong assumption must be made by the experimenter under this framework: His/her degree of uncertainty about model adequacy. A Bayesian approach is proposed to deal with this assumption. A new optimality criterion is proposed, and numerical algorithms are provided to calculate this new class of optimal robust designs. Several examples illustrate the results.

C0490: **Dynamic design of experiments for function-on-function linear models**
Presenter: **Caterina May**, Kings College London, United Kingdom

The optimal design of experiments in the context of functional data has been explored very little until now. Linear models are considered where both the response and one or more factors are continuous functions, for example, of the time. The goal is to find the optimal dynamic experimental conditions to estimate precisely the functional coefficients of the model. After establishing a suitable estimator and obtaining its variance-covariance matrix, the definition of optimal design criteria is extended to this functional context. A-optimal and D-optimal functional designs are then computed in practice, using the choice of suitable bases of functions to represent the data. The efficiency of estimators obtained under different choices is then compared.

C0831: **Optimized recovery sampling to test for missing not at random**

Presenter: **Robin Mitra**, UCL, United Kingdom

Missing data can lead to inefficiencies and biases in analyses, in particular when data are missing, not at random (MNAR). It is thus vital to understand and correctly identify the missing data mechanism. Recovering missing values through a follow-up sample allows researchers to conduct hypothesis tests for MNAR, which is not possible when using only the original incomplete data. Investigating how the properties of these tests are affected by the follow-up sample design is not explored in the literature. The results provide comprehensive insight into the properties of one such test based on the commonly used selection model framework. Conditions are determined for recovery samples that allow the test to be applied appropriately and effectively, i.e. with known type I error rates and optimized with respect to power. An integrated framework is thus provided for testing the presence of MNAR and designing follow-up samples in an efficient, cost-effective way. The methodology's performance is evaluated through simulation studies and on a real data sample.

CO237 **Room K2.31 (Nash Lec. Theatre)** **HIGH-DIMENSIONAL STATISTICS FOR GENOMICS AND BIOMEDICINE** **Chair: Mayetri Gupta**

C0566: **Statistical and machine learning models from resolving metabolomics spectra**

Presenter: **Vinny Davies**, University of Glasgow, United Kingdom

Co-authors: Nikolaos Terzis, Andrew Elliott, Ronan Daly, Joe Wandy

Untargeted metabolomics experiments aim to identify the small molecules that make up a particular sample, e.g., blood or urine. The sample is put through a mass spectrometer, which performs multiple scans of different types, giving us a large amount of data which can be used to estimate the spectra which are needed to identify the metabolites. In a particular type of experiment, known as data-independent acquisition, large amounts of data representing metabolite fingerprints, known as fragments, are collected, which can be used to reconstruct the metabolite spectra within the samples. The challenge, however, with this method is that it is not clear to which metabolite a given fragment belongs. Data related to the abundance of the metabolites is used within the samples to link the measured fragments back to the metabolite from which they originated. A modelling framework is created that uses a large number of high dimensional regression models to create predictions for the metabolite spectra. The statistical framework is introduced, and a multitude of modelling methods used are described to tackle this difficult and very noisy task.

C0993: **A Bayesian functional factor model for high-dimensional molecular curves**

Presenter: **Helene Ruffieux**, Univerisity of Cambridge, United Kingdom

Co-authors: Salima Jaoua, Daniel Temko

The increasing availability of longitudinal measurements on gene products is set to improve the understanding of the molecular processes underlying disease risk and progression. However, methods for modelling coordinated patterns of temporal variation are currently lacking. A Bayesian approach for representing high-dimensional curves is proposed, combining latent factor modelling and functional principal component analysis (FPCA). This approach captures correlations across variables (gene products) and time by positing that a subset of variables contributes to a small number of FPCA expansions (representing latent disease processes) through variable-specific loadings. Subject variability is modelled using a few functional principal components, each characterised by a smoothly varying temporal function. A variational inference algorithm is introduced, with analytical updates, coupling efficiency and principled parameter uncertainty quantification. Extensive numerical experiments illustrate the ability of the approach to (i) accurately estimate variable-specific loadings, FPCA latent functions and subject-specific component scores and (ii) scale to realistic molecular data sizes. This framework should help disentangle disease heterogeneity by clarifying how gene pathways coordinate over time and predicting molecular trajectories at the subject level, thus facilitating targeted interventions and personalized treatments.

C1073: **Challenges in estimation of small genetic effects in large-scale population cohorts**

Presenter: **Ava Khamseh**, University of Edinburgh, United Kingdom

Co-authors: Sjoerd Beentjes, Chris Ponting, Olivier Labayle, Mark van der Laan, Kelsey Tetley-Campbell, Joshua Slaughter

A key aim in genomic medicine is the identification of likely causal DNA variants altering human traits or diseases. Such causal genetic support is crucial for efficient drug discovery because it is estimated to double the success rate of drugs in clinical development. The concrete challenge in this area is that while the data size is very large (100K-1M samples), the effect sizes of individual DNA variants on disease outcomes are expected to be relatively small. This means the slightest degree of bias due to model misspecification can result in biased estimates that may be falsely prioritised for costly experimental verification. The problem is exacerbated when attempting to estimate higher-order DNA variant interactions on disease. At the same time, millions of estimations are performed to probe the genome. This implies that the scalability of the algorithms used is essential. TarGene is introduced as a methodology, pipeline, and software to reproducibly and reliably estimate genetic effect sizes on traits or diseases through various semi-parametric efficient techniques. Although asymptotic properties are equal, these estimators may perform differently in finite samples, necessitating careful comparisons. Open challenges remain, such as (i) accounting for population stratification (beyond PCA) and (ii) correlated causal variants near a variant of interest.

C1665: **ZINBGT: Exploratory data analysis of transcriptomic expression using mixture models**

Presenter: **Toby Kettlewell**, University of Glasgow, United Kingdom

Co-authors: Yiyi Cheng, Thomas Otto, Vincent Macaulay, Mayetri Gupta

Single-cell RNA sequencing (scRNA-seq) provides data on the signals associated with protein production within individual cells. This allows for the discovery of novel cell types, inference of cell trajectories, and fine-grained comparisons of different tissues. The analysis of scRNA-seq data uses a pipeline of methods, but benchmarking is currently unable to establish which methods are best for a given dataset. Because the conclusions drawn from an analysis depend on the choice of method, novel forms of exploratory data analysis are needed to investigate how datasets differ and the circumstances in which a given method is likely to perform best. Any such method needs to run quickly and provide easily interpretable visualisations. A family of mixture distributions on count data will be introduced, which capture the salient aspects of gene expression with the associated parameters acting as summary statistics of each gene. A novel variant of a 2d histogram will be proposed, allowing efficient exploration and comparison of large, high-dimensional datasets, while problematic genes are highlighted using a combination of a distance between model and data with bootstrapping. Human immune cells will be explored in terms of gene expression, and comparison with simulations will reveal differences that could compromise benchmarking.

CO352 **Room K2.40** **HIGH-DIMENSIONAL STATISTICS WITH NUISANCE PARAMETERS** **Chair: Tengyao Wang**

C1296: **Inference on the significance of modalities in multimodal generalized linear models**

Presenter: **Quefeng Li**, University of North Carolina - Chapel Hill, United States

Co-authors: Wanting Jin

Multimodal statistical models have gained much attention in recent years, yet there is a lack of rigorous statistical inference tools for inferring the significance of a single modality within a multimodal model. This inference problem is particularly challenging in high-dimensional multimodal models. In the context of high-dimensional multimodal generalized linear models, a novel entropy-based metric called the expected relative entropy (ERE) is proposed to quantify the information gain of one modality in addition to all other modalities in the model. A deviance-based statistic is then proposed to estimate the ERE. The deviance-based statistic is proven consistent with the ERE, and its asymptotic distribution is derived, which enables the calculation of confidence intervals and p-values to assess the significance of the ERE for a given modality. The empirical performance of the proposed inference tool is numerically evaluated on various high-dimensional multimodal generalized linear models, and its good performance is demonstrated in these models. The method is also applied to a multimodal neuroimaging dataset to demonstrate its capability to infer the significance of imaging modalities, which is crucial for neuroscience studies.

C1644: **Inference of transition time in steady-state variations in smFRET via a Wasserstein distance approach**
Presenter: **Fengnan Gao**, University College Dublin, Ireland

Many biological molecules respond to external stimuli that can cause their conformational states to shift from one steady state to another. To study steady-state transitions in single-molecule fluorescence resonance energy transfer, a novel methodology is introduced, called WAVE (Wasserstein distance Analysis in steady-state Variations in smFRET) to detect and locate non-equilibrium transition positions in FRET trajectories. The method first utilizes a combined STaSI-HMM (stepwise transitions with state inference hidden Markov model) algorithm to convert the original FRET trajectories into discretized trajectories. Maximum Wasserstein distance analysis is then applied to differentiate the FRET state compositions of the fitting trajectories before and after the non-equilibrium transition. This methodology allows observing changes in experimental conditions in chromophore-tagged biomolecules or vice versa. Statistical properties, such as consistency and convergence rates, of the methodology are investigated under mild conditions.

C1648: **Residual permutation test for high-dimensional regression coefficient testing**
Presenter: **Yuhao Wang**, Tsinghua University and Shanghai Qi Zhi Institute, China
Co-authors: Kaiyue Wen, Tengyao Wang

The problem of testing whether a single coefficient is equal to zero in fixed-design linear models under a moderately high-dimensional regime is considered, where the dimension of covariates p is allowed to be in the same order of magnitude as sample size n. In this regime, to achieve finite-population validity, existing methods usually require strong distributional assumptions on the noise vector (such as Gaussian or rotationally invariant), which limits their applications in practice. A new method, called residual permutation test (RPT) is proposed, which is constructed by projecting the regression residuals onto the space orthogonal to the union of the column spaces of the original and permuted design matrices. RPT can be proved to achieve finite-population size validity under fixed design with just exchangeable noises whenever $p < n/2$. Moreover, RPT is shown to be asymptotically powerful for heavy-tailed noises with bounded $(1+t)$-th order moment when the true coefficient is at least of order $n^{-t/(1+t)}$ for $t \in [0,1]$. It is further proven that this signal size requirement is essentially rate-optimal in the minimax sense. Numerical studies confirm that RPT performs well in a wide range of simulation settings with normal and heavy-tailed noise distributions.

C1411: **Bias correction in factor-augmented regression models**
Presenter: **Peiyun Jiang**, Tokyo Metropolitan University, Japan
Co-authors: Yoshimasa Uematsu, Takashi Yamagata

A new estimation procedure is introduced for factor-augmented regression models to eliminate the bias of the ordinary least squares (OLS) estimators. The asymptotic properties of the OLS estimators are first derived, allowing for weak factors. It is demonstrated that the bias inherent in the common estimation approach is driven by the dependence between latent factors and observed regressors, the divergence rates of the signal eigenvalues, and the choice of the rotation matrix. An interesting result is that the commonly used rotation matrix generates the largest bias in the factor-augmented regression estimators, whereas an alternative rotation matrix can reduce the bias to some extent. To address these issues, a novel estimation procedure is proposed that projects out the observed regressors during the principal component (PC) estimation step, thereby mitigating the correlation between latent factors and regressors. The procedure induces exactly zero bias under weaker conditions than those required by existing methods. Monte Carlo simulations confirm that while common approaches suffer from significant bias and severe size distortion in testing, the proposed method demonstrates satisfactory accuracy in estimation and optimal testing performance in terms of size and power.

CO176 Room K2.41 EXTEME RISKS ANALYSIS AND REAL LIFE APPLICATIONS Chair: Maud Thomas

C0689: **Assessing extreme risk using stochastic simulation of extremes**
Presenter: **Nisrine Madhar**, LPSM, France
Co-authors: Juliette Legrand, Maud Thomas

Risk management is particularly concerned with extreme events, but analysing these events is often hindered by the scarcity of data, especially in a multivariate context. This data scarcity complicates risk management efforts. Various tools can assess the risk posed by extreme events, even under extraordinary circumstances. The evaluation of univariate risk is studied for a given risk factor using metrics that account for its asymptotic dependence on other risk factors. Data availability is crucial, particularly for extreme events where it is often limited by the nature of the phenomenon itself, making estimation challenging. To address this issue, two non-parametric simulation algorithms based on multivariate extreme theory are developed. These algorithms aim to extend a sample of extremes jointly and conditionally for asymptotically dependent variables using stochastic simulation and multivariate generalized Pareto distributions. The approach is illustrated with numerical analyses of both simulated and real data to assess the accuracy of extreme risk metric estimations.

C0794: **Modeling moderate and extreme rainfall at high spatiotemporal resolution**
Presenter: **Nicolas Meyer**, Universite de Montpellier, France

Flood risk analysis in an urban environment requires a precise understanding of rainfall events. The specific case of the Montpellier region is presented, in which a network of rain gauges from the Montpellier urban observatory enables carrying out an initial study based on data for the period 2019-2022. As rainfall events are rare (but intense) in this region, only a few data are available. Therefore, the existing dataset is enriched with a stochastic rainfall generator. The main steps are presented in the construction of this generator: simulation of marginals according to the extended generalized Pareto distribution and simulation of the spatiotemporal dependence structure via a Brown-Resnick or r-Pareto process, taking into account physical constraints such as wind via advection. This leads to a full model for rainfall events for which a statistical analysis is proposed.

C1007: **Parametric reinsurance for extreme claims**
Presenter: **Olivier Lopez**, Ensae IP Paris, France

Parametric (or index-based) insurance products are increasingly used to cover emerging risks, especially the upper segments not covered by traditional insurance. The idea is not to cover directly the risk but to compute the compensation based on the value of a measurable index just after the claim. From a risk management perspective, controlling this index is easier, allowing insurability. The particular case is discussed, where the loss distribution is the heavy tail (Pareto tail) and may not even have an expectation. Adapting a utility framework, it is shown how to conceive optimal reinsurance products based on the parametric approach.

C1081: **Analyzing the dynamics of extreme events with marked point processes**
Presenter: **Antoine Heranval**, INRAE, France
Co-authors: Thomas Opitz, Denis Allard

The aim is to merge probabilistic methods for marked point processes, extreme events, and spatial graphs in order to propose a new framework for better understanding spatial and temporal interactions between different types of extreme events in the climate system. Each extreme event (whether temperature, precipitation, wind speed, atmospheric pressure, etc.) for a specific climate variable and region will be represented by a dot indicating the time of its occurrence, as well as by one or more markers describing key characteristics of the episode, such as maximum, spatial extent, duration, accumulation, etc. First- and second-order statistical characteristics and second-order statistical features of point processes, such as the intensity function and various pair and mark correlation functions, will be adapted to study the multiscale structures of the set of marked points. New classes of models are also developed for marked point processes, defined on a spatial graph representing the geographical proximity between regions, in order to simulate new realizations of the set of such events.

CO340 Room S0.03 PROBABILISTIC PREDICTION OF COMPLEX DATA **Chair: Matteo Fontana**

C0359: **Nonparametric multiple-output center-outward quantile regression**
Presenter: **Alberto Gonzalez Sanz**, Columbia University, United States

The problem of nonparametric multiple-output quantile regression is addressed based on the novel concept of multivariate center-outward quantiles introduced in past studies. To obtain conditional quantile regions and contours, the conditional center-outward quantiles are defined. A new cyclically monotone interpolation, with non-necessarily constant weights, is proposed to define them. This method is completely nonparametric and produces interpretable empirical regions/contours that converge in probability with its population counterpart. Some real and synthetic examples are included, showing the adaptation and applicability of the method, in particular its ability to catch the heteroskedasticity and the trend of the data.

C0453: **Generative machine learning methods for multivariate ensemble postprocessing**
Presenter: **Sebastian Lerch**, Karlsruhe Institute of Technology, Germany

Ensemble weather forecasts based on multiple runs of numerical weather prediction models show systematic errors and require postprocessing. Accurately modeling multivariate dependencies is crucial in many practical applications, and various approaches to multivariate postprocessing have been proposed where ensemble predictions are first postprocessed separately in each margin, and multivariate dependencies are then restored via copulas. These two-step methods share common key limitations, particularly the difficulty of including additional predictors in modelling the dependencies. The novel multivariate postprocessing method is based on generative machine learning to address these challenges. In this new class of nonparametric data-driven distributional regression models, samples from the multivariate forecast distribution are directly obtained as a generative neural network output. The generative model is trained by optimizing a proper scoring rule, which measures the discrepancy between the generated and observed data, conditional on exogenous input variables. The method does not require parametric assumptions on univariate distributions or multivariate dependencies and allows for incorporating arbitrary predictors. In two case studies on multivariate temperature and wind speed forecasting at weather stations over Germany, our generative model shows significant improvements over state-of-the-art methods and particularly improves the representation of spatial dependencies.

C0925: **Angular combining of forecasts of probability distributions: Applications and developments**
Presenter: **James Taylor**, University of Oxford, United Kingdom
Co-authors: Xiaochun Meng

To enable a pragmatic synthesis of the information available in different forecasts of a probability distribution, forecast combining can be used. The combination has often been applied to the probability predictions of the distributional forecasts. However, it has been suggested that this will tend to deliver overdispersed distributional forecasts, prompting the combination to be applied instead to the quantile predictions. The probability combining approach can be viewed as vertical combining of the probability distributions, and the quantile combining approach is viewed as horizontal combining. The proposal is to combine at an angle between the extreme cases of vertical and horizontal combining, with the angle optimized using a proper scoring rule. For implementation, a pragmatic numerical approach and a simulation algorithm are provided. Among the theoretical results, it is shown that, as with vertical and horizontal averaging, angular averaging results in a distribution with a mean equal to the average of the means of the distributions that are being combined. It is also shown that angular averaging produces a distribution with lower variance than vertical averaging and, under certain assumptions, greater variance than horizontal averaging. An empirical illustration of angular combining using several different applications is provided. Finally, potential extensions and generalizations are described.

C0941: **Large language model validity via enhanced conformal prediction methods**
Presenter: **John Cherian**, Stanford University, United States
Co-authors: Isaac Gibbs, Emmanuel Candes

New conformal inference methods are developed to obtain validity guarantees on the output of large language models (LLMs). Prior work in conformal language modeling identifies a subset of the text that satisfies a high-probability guarantee of correctness. These methods work by filtering claims from the LLM's original response if a scoring function evaluated on the claim fails to exceed a threshold calibrated via split conformal prediction. Existing methods in this area suffer from two deficiencies. First, the guarantee stated is not conditionally valid. The trustworthiness of the filtering step may vary based on the topic of the response. Second, because the scoring function is imperfect, the filtering step can remove many valuable and accurate claims. Both of these challenges are addressed via two new conformal methods. First, the conditional conformal procedure is generalized to adaptively issue weaker guarantees when required to preserve the utility of the output. Second, it is shown how to systematically improve the quality of the scoring function via a novel algorithm for differentiating through the conditional conformal procedure. The efficacy of the approach is demonstrated on both synthetic and real-world datasets.

CO045 Room S0.11 SPATIAL INFERENCE FOR fMRI ANALYSIS **Chair: Armin Schwartzman**

C1647: **Quantifying the spatial uncertainty of excursion sets**
Presenter: **Armin Schwartzman**, University of California, San Diego, United States
Co-authors: Junting Ren, Fabian Telschow

A central problem in image analysis, particularly in brain mapping, is locating the important effects spatially. The standard solution has been to treat it as a large-scale multiple-testing problem. However, this approach assumes signal sparsity and does not provide a measure of spatial uncertainty. It is proposed to directly address the question of where the important effects are by estimating the excursion set where the signal is greater than a threshold. To assess uncertainty, spatial confidence regions are constructed, given as nested sets that spatially bound the true excursion set with a given probability. It is shown that confidence regions with simultaneous control over all excursion thresholds can be obtained by thresholding standard simultaneous confidence bands. This approach is developed for excursion sets of the mean function in a signal-plus-noise model, including coefficients in pointwise regression models, as those used in task fMRI analysis.

C1629: **Simultaneous confidence regions of excursion sets**
Presenter: **Fabian Telschow**, Humboldt University zu Berlin, Germany
Co-authors: Armin Schwartzman, Junting Ren

Asymptotic statistical inference in the space of bounded functions endowed with the supremum's norm over an arbitrary metric space S is studied using simultaneous confidence regions of excursion (SCoRE) sets. These sets simultaneously quantify the uncertainty of several lower and upper excursion sets of a target function. Their connection is investigated to multiple hypothesis tests controlling the familywise error rate (FWER) in the strong sense, and it is shown that they grant a unifying perspective on statistical inference tools such as simultaneous confidence bands, quantification of uncertainties in level set estimation, for example, CoPE sets, and multiple hypothesis testing over S, for example, finding relevant differences or regions of equivalence within S. Additionally, the abstract setting allows refining and reducing the assumptions in recent articles on

CoPE sets and relevance and equivalence testing using the supremum's norm as well as propose novel relevance and equivalence tests that control the FWER in the strong sense for Banach spaced data.

C1626: **Spatial confidence regions for combinations of excursion sets in image analysis**
Presenter: **Thomas Maullin-Sapey**, University of Bristol, United Kingdom

The comparison of random fields obtained across the same spatial region but under different study conditions is essential to a wide range of disciplines such as neuroimaging, climatology and cosmology. While much literature has documented the geometric and topological properties of excursion sets of random fields, surprisingly, little has been published concerning how such excursion sets may be compared to one another. The aim is to describe how confidence regions (CRs) may be used to characterize the spatial uncertainty of the intersection and union of several excursion sets obtained under different study conditions. These regions are of natural interest as they directly correspond to questions of the form "Where do all random fields exceed a predetermined threshold?" and "Where does at least one random field exceed a predetermined threshold?". The general method is first described, and then an illustration of its use is provided in assessing spatial uncertainty in fMRI analyses of human brain activity.

C1630: **Transforming heavy tailed data improves the power and validity of inference**
Presenter: **Samuel Davenport**, University of California, San Diego, United States

The purpose is to demonstrate that transforming heavy-tailed data leads to improvements in both the power and validity of inference, with a particular focus on applications in neuroimaging in which the noise can be very heavy-tailed. Validity is improved because the rate of convergence of the CLT is accelerated, leaving valid p-values under the null. Instead, a power gain occurs because the transformation can increase the value of Cohens d. It is shown that transformations can also be combined with sign-flipping to infer the null distribution of the transformed data, ensuring validity and providing a power boost. The approach is validated using gold standard resting state fMRI null simulations and task fMRI datasets used from the Human Connectome Project to illustrate the improvements in power.

CO229 Room S0.12 EXTREMES AND LEVY PROCESSES Chair: Ana Ferreira

C0226: **General graphical models for stable processes**
Presenter: **Florian Brueck**, University of Geneva, Switzerland

General graphical models are introduced for multivariate Levy processes with stable margins. First, Huesler-Reiss Levy processes are defined in terms of a Levy measure, which is derived from the exponent measure of a Huesler-Reiss distribution. The Huesler-Reiss Levy processes can be equipped with arbitrary alpha-stable margins, and their conditional independence structure is encoded in a single d-times-d matrix. Furthermore, it is shown how the conditional independence structure of a Huesler-Reiss Levy process may be estimated, and a simulation scheme is provided to generate trajectories of the corresponding paths.

C0346: **On asymptotic independence in higher dimensions**
Presenter: **Vicky Fasen-Hartmann**, Karlsruhe Institute of Technology, Germany
Co-authors: Bikramjit Das

In the study of extremes, the presence of asymptotic independence signifies that extreme events across multiple variables are probably less likely to occur together. Although well-understood in a bivariate context, the concept remains relatively unexplored when addressing the nuances of joint occurrence of extremes in higher dimensions. A notion of mutual asymptotic independence is proposed to capture the behavior of joint extremes in dimensions larger than two and contrast it with the classical notion of (pairwise) asymptotic independence. Furthermore, k-wise asymptotic independence is defined as a relationship between pairwise and mutual asymptotic independence. The concepts are compared using examples from Archimedean, Gaussian, and Marshall-Olkin copulas, among others. Notably, for the popular Gaussian copula, explicit conditions are provided on the correlation matrix for mutual asymptotic independence to hold; moreover, exact tail orders are computed for various tail events.

C0349: **Assessing causality in the tails: Measurement and testing**
Presenter: **Bikramjit Das**, Singapore University of Technology and Design, Singapore
Co-authors: Xiangyu Liu

A measure of extreme tail association (ETA) is defined between two variables, which is asymmetric, easily computable in sample data, and consistent for the population measure. The asymptotic normality of the sample measure is exhibited under mild conditions on the underlying distribution. A test for bivariate tail asymmetry is also proposed, and asymptotic distributions are computed under both null and alternative hypotheses. Confidence regions for real data are computed using a multiplier bootstrap method. The measures and tests established allow inferring causality relationship between two variables. Data from movement in cryptocurrency prices is used to exhibit the the tools developed.

C1038: **Weak subordination of multivariate Levy processes**
Presenter: **Boris Buchmann**, Australian National University, Australia

Subordination is the operation that evaluates a Levy process at a subordinator, giving rise to a path-wise construction of a "time-changed" process. In probability semigroups, subordination was applied to create the variance gamma process, which is prominently used in financial modelling. However, subordination may not produce a levy process unless the subordinate has independent components or the subordinate has indistinguishable components. A new operation known as weak subordination is introduced that always produces a Levy process by assigning the distribution of the subordinate conditional on the value of the subordinate, which matches traditional subordination in law in the cases above. Weak subordination is applied to extend the class of variance-generalized gamma convolutions and to construct the weak variance-alpha-gamma process. The latter process exhibits a wider range of dependence than using traditional subordination.

CO007 Room S0.13 RECENT TOPICS IN BIOSTATISTICS Chair: Kathrin Moellenhoff

C0348: **Overcoming model uncertainty: How equivalence tests can benefit from model averaging**
Presenter: **Niklas Hagemann**, University of Cologne, Germany
Co-authors: Kathrin Moellenhoff

A common problem in numerous research areas, particularly in clinical trials, is to test whether the effect of an explanatory variable on an outcome variable is equivalent across different groups. In practice, these tests are frequently used to compare the effect between patient groups, e.g., based on gender, age, or treatments. Equivalence is usually assessed by testing whether the difference between the groups does not exceed a pre-specified equivalence threshold. Such tests are often based on the distance between two parametric models. These approaches have one thing in common: They are based on the assumption that the true underlying regression model is known. A flexible extension of such methodology is proposed that uses model averaging in order to overcome this assumption and make the test applicable under model uncertainty. Model averaging is introduced based on smooth AIC weights. In order to ensure numerical stability, a testing procedure is proposed which makes use of the duality between confidence intervals and hypothesis testing. The validity of the approach is demonstrated by a simulation study. A time-response case study demonstrates the practical relevance of the approach.

C0557: **Accounting for delayed responses in group sequential clinical trial designs**
Presenter: **Carolin Herrmann**, Heinrich Heine University Duesseldorf, Germany

Group sequential clinical trial designs are applied frequently and come with the advantage of potentially earlier decisions and reduced sample sizes.

In their interim analyses, decisions are made on whether the trial is terminated early with a potential rejection of the null hypothesis or whether the trial is continued with another stage. Often, there occurs a time gap between enrolling a patient into a trial and measuring the respective outcome. This potential delay in the availability of the outcome measurement can affect the statistical analyses, especially when it comes to interim analyses. Some statistical methods that account for those delayed responses have already been proposed. The idea is often twofold: First, a decision is made on an irreversible recruitment stop, followed by a hypothesis test in case of stopping. Modifications of this idea are proposed. Moreover, results from a simulation study are presented where different options of accounting for delayed responses are compared with the option of not accounting for them. Results are discussed with respect to power and sample size. There will be a special focus on potential performance differences as well as the influence of the amount of pipeline data and interim information.

C0564: **Enhancing time-to-event prediction with high-dimensional omics data using exclusive Lasso regularization**
Presenter: **Dayasri Ravi**, Technical University Dortmund, Germany
Co-authors: Andreas Groll

The integration of high-dimensional omics data into survival prediction models has gained significant attention due to the growing availability of these datasets. Traditionally, a Cox regression model is employed, concatenating various omics data types linearly. Given that much of the omics data may be redundant or irrelevant, feature selection through penalization is often necessary. A notable characteristic of these datasets is their organization into blocks of distinct data types, such as methylation and clinical data, which requires selecting a subset of variables from each group due to high intra-group correlations. The proposal is to utilize exclusive Lasso regularization in place of the standard Lasso penalty. Exclusive Lasso promotes intra-group sparsity via the L_1-norm while encouraging inter-group selection through the L_2-norm, ensuring the selection of at least one variable per group. A significant challenge arises from the non-differentiability of the L_1-norm within groups, which is addressed by transforming this norm using quadratic and sigmoid function approximations to achieve differentiability. This transformation facilitates the use of a straightforward Newton-based approach to solve the intricate optimization problem. The methodology is applied to real-life cancer datasets, demonstrating enhanced survival prediction performance compared to the conventional Lasso-penalized Cox regression model.

C0929: **Recent challenges and biostatistical approaches in polygenic risk score modelling**
Presenter: **Christian Staerk**, IUF-Leibniz Research Institute for Environmental Medicine, Germany
Co-authors: Hannah Klinkhammer, Carlo Maj, Andreas Mayr

Polygenic risk scores (PRS) quantify the genetic predisposition for various traits and clinical outcomes based on genotype data. The aim is to address recent challenges in PRS modelling from a biostatistical perspective. In particular, the focus is on three important challenges: first, training PRS models on high-dimensional and large-scale genotype data with hundreds of thousands of genetic variants and individuals requires scalable and interpretable statistical learning methods. Second, the transferability of PRS models to diverse populations with different ancestries is often limited, as models are typically trained on cohorts predominantly of European ancestry. Third, the evaluation of the prediction accuracy of PRS models is complicated by different and conflicting definitions of the commonly used R-squared measure on test data. To address these challenges, recent statistical learning approaches for fitting PRS models are presented on individual-level genotype data from the UK Biobank, including scalable boosting and causal inference methods. Furthermore, open problems for future research are highlighted to enhance the precision and integration of PRS models for personalized medicine.

CO134 Room Safra Lec. Theatre ADVANCES IN FUNCTIONAL AND OBJECT ORIENTED DATA ANALYSIS Chair: Alessia Pini

C0574: **Testing missing completely at random for partially observed functional data**
Presenter: **Maximilian Ofner**, Graz University of Technology, Austria
Co-authors: Siegfried Hoermann, David Kraus, Dominik Liebl

The statistical analysis of incompletely observed functional data, referred to as partially observed functional data, has gained considerable attention recently. Corresponding data can result from issues like malfunctioning measuring devices that fail to record the signals over certain periods. Current methods for analysing such data typically rely on a missing completely at random (MCAR) assumption, meaning the missingness mechanism is independent of the data itself. However, limited focus has been given to verifying this assumption. This gap is addressed by proposing novel procedures to test MCAR for different types of random functions. In addition, asymptotic distributions are established, and the consistency of the tests is discussed under various alternatives. The finite sample properties of the methods are then evaluated through a simulation study. Finally, a real-data application illustrates the practical utility of the methodology, highlighting its potential not only for justifying statistical methods but also for identifying the cause of missingness.

C1226: **Statistical analysis of trajectories of 3-dimensional object orientations**
Presenter: **Aymeric Stamm**, CNRS, France
Co-authors: Lise Bellanger, Margot Bornet, Klervi Le Gall, Nadia Negab, Manon Simonot

It is becoming increasingly frequent to measure a collection of trajectories of 3-dimensional object orientations in various fields such as robotics, tracking in video, health, etc. Orientation is mathematically expressed as a 3D rotation matrix. Various other representations can be used: Euler angles, axis-angle pairs and unit quaternions. The latter representation is often preferred as the most compact, avoiding the gimbal lock issue. Trajectories of 3D object orientations are, therefore, typically treated as unit quaternion time series. The aim is to present a comprehensive framework for the analysis of samples of unit quaternion time series, which unlocks the use of traditional statistical methods such as summary statistics, dimensionality reduction, supervised and unsupervised classification, prediction, etc. Specifically, unit quaternions belong to the special unitary group SU(2), which is a Lie group. As such, it is both a group and a differentiable Riemannian manifold. The latter property guarantees the existence of tangent spaces at each point of the manifold, onto which classical Euclidean geometry of vector spaces applies. Metrics are introduced with geometric invariance on tangent spaces and demonstrate the use of the framework to study and compare individual gait patterns.

C1421: **Metric statistics for geometric graphs with application to cardiac fibrosis**
Presenter: **Anna Calissano**, Imperial College London, United Kingdom
Co-authors: Arstanbek Okenov, Alexander Panfilov, Timur Nezlobinsky

Geometric graphs (or spatial networks) are graphs in which the nodes are embedded as points in Euclidean space, and the edges encode certain geometrical properties of an underlying object. Such graphs can describe the skelethon of a shape, often depicted in an image. The purpose is to describe how the fused Gromov-Wasserstain metric can be used to compare different geometric graphs with different numbers of nodes and edges. The metric is adjusted to leverage rotations between the graphs, and we use the metric to perform clustering and testing procedures. As a motivational case study, the geometry and the topology of cardiac fibrosis are studied in the human heart. Cardiac fibrosis is the pathological increase of fibroblast in the heart. The data consists of histopathological images of human hearts showing cardiac fibrosis. Each image representing the fibroblast is turned into a bidimensional spatial network using skelethonization. Via clustering procedure and metric-based ANOVA, the question of how cardiac fibrosis is structured in terms of its geometrical and topological characteristics is addressed, along with the existing variability within a single patient and across patients.

C0599: **Adaptive functional regression for locally heterogeneous spectroscopic data**
Presenter: **Alessandro Casa**, Free University of Bozen-Bolzano, Italy
Co-authors: Federico Ferraccioli , Marco Stefanucci

Mid-infrared spectroscopy is a valuable tool for collecting vast amounts of data quickly and in a relatively cheap way. These data provide a rich reservoir of information about analyzed samples, which have been used in various scientific fields. However, spectroscopic data present some serious challenges from a statistical viewpoint. In fact, each spectrum is a curve typically consisting of more than 1000 absorbance values measured across different wavelengths, with a large portion of the spectral domain known to be highly noisy and not containing relevant information for the scopes of the analyses. Nonetheless, variable selection procedures are often hindered by redundancies and complex relationships among wavelengths. Additionally, the signal is highly heterogeneous across the spectral domain, often limiting the use of standard functional data analysis tools. To address these challenges, we introduce an adaptive scalar-on-function regression tool that preserves the functional nature of the data while accommodating varying degrees of smoothness and inhomogeneous signals. Our framework allows us to include scalar covariates and to model both Gaussian and non-Gaussian responses, expanding its use to classification tasks and regression for count data. Furthermore, we propose a bootstrap-based inferential procedure to identify the spectral regions most influential in predicting response variables.

CO122 Room BH (S) 1.01 Lec. Theatre 1 MODELLING NON-STATIONARITY AND STRUCTURAL CHANGE Chair: Liudas Giraitis

C0814: **Estimation of random cycles in persistent time series**
Presenter: **Liudas Giraitis**, Queen Mary University of London, United Kingdom
Co-authors: Karim Abadir , Natalia Bailey, Walter Distaso

A number of economic, financial, and climatic time series exhibit persistent cycles which are characterized by dependence patterns and peaks in the spectrum. A class of semiparametric cyclical-memory processes is introduced, which enable the modelling of random cyclical patterns in stationary and non-stationary time series. A theoretical background and asymptotic estimation theory are developed for the frequency of a cycle represented by the location of a peak in the spectrum. The estimation procedure is easy to implement and allows for constructing narrow confidence intervals around the location point. Monte Carlo simulations confirm the estimator's good finite sample performance. The method is illustrated with three empirical applications. Quasi- periodic cycles are uncovered in macroeconomic series, both nominal and real (US nominal GDP and real industrial production), and CO2 concentration levels.

C0702: **Asymptotic theory for constant step size stochastic gradient descent**
Presenter: **Stefan Richter**, HHU Duesseldorf, Germany
Co-authors: Wei Biao Wu, Zhipeng Lou, Jiaqi Li

A novel approach is presented to understanding the behavior of stochastic gradient descent (SGD) with constant step size by interpreting its evolution as a Markov chain. Unlike previous studies that rely on the Wasserstein distance, this approach leverages the functional dependence measure, and the geometric-moment contraction (GMC) property is explored to capture the general asymptotic behavior of SGD in a more refined way. In particular, the approach allows SGD iterates to be non-stationary but asymptotically stationary over time, providing quenched versions of the central limit theorem and invariance principle valid for averaged SGD with any given starting point. A Richardson-Romberg extrapolation is subsequently defined with an improved bias representation to bring the estimates closer to the global optimum. The existence of a stationary solution is established for the derivative SGD process under mild conditions, enhancing the understanding of the entire SGD procedure. Lastly, an efficient online method is proposed for estimating the long-run variance of SGD solutions.

C0900: **Regression with lagged variables in heterogeneous environment**
Presenter: **Yufei Li**, Kings College London, United Kingdom
Co-authors: Liudas Giraitis, George Kapetanios

The recent work on regression modelling that permits for general heterogeneity is extended to allow for lagged dependent variables. The purpose is to explore to what extent the generality of the setting, the simplicity of assumptions, and the ease of computation of standard errors can be preserved. Theoretical properties of regression estimation and inference is accompanied by Monte Carlo experiments and an empirical application.

C0629: **Common trends and long-run identification in nonlinear structural VARs**
Presenter: **James Duffy**, Oxford, United Kingdom
Co-authors: Sophocles Mavroeidis

While it is widely recognised that linear (structural) VARs may omit important features of economic time series, the use of nonlinear SVARs has to date been almost entirely confined to the modelling of stationary time series because of a lack of understanding as to how common stochastic trends may be accommodated within nonlinear VAR models. This has, unfortunately, circumscribed the range of series to which such models can be applied and/or required that these series be first transformed to stationarity, a potential source of misspecification, and prevented the use of long-run identifying restrictions in these models. To address these problems, a flexible class of additively time-separable nonlinear SVARs is developed, which subsume models with threshold-type endogenous regime switching, both of the piecewise linear and smooth transition varieties. The Granger-Johansen representation theorem is extended to this class of models, obtaining conditions that specialize exactly to the usual ones when the model is linear. It is further shown that, as a corollary, these models are capable of supporting the same kinds of long-run identifying restrictions as are available in linear cointegrated SVARs.

CO231 Room BH (SE) 1.02 PREDICTION AND CLUSTERING UNDER BAYESIAN MIXTURE MODELS Chair: Raffaele Argiento

C0302: **Matrix-variate priors for flexible mixture modelling of grouped data**
Presenter: **Andrea Cremaschi**, IE University, Spain
Co-authors: Beatrice Franzolini

In the last two decades, the Bayesian nonparametric literature witnessed significant progress in the study of novel dependent prior distributions beyond standard univariate species sampling processes. These dependent priors are developed under partial exchangeability assumptions and capture dependencies in heterogeneous data settings with grouped data. A notable example is the celebrated hierarchical Dirichlet process. However, with a few exceptions, less attention has been given to finite-dimensional dependent mixture models and dependent mixtures with a random number of components. A novel class of dependent priors is proposed for mixture modelling based on finite-dimensional matrix-variate distributions for the weights of the mixture. Specifically, the matrix-variate Dirichlet distribution is employed as a joint prior for the weights of the multi-group mixture. The distributional properties of the matrix-variate Dirichlet distribution ensure standard assumptions for the weights of each mixture while inducing dependence and appropriate borrowing of information across groups. The approach goes beyond standard univariate weights, allowing for varying levels of description of the data features and accommodating group-specific kernels. This enables flexible modelling of different data types and various ways in which the information can be shared across groups. The proposed model is widely applicable and yields interpretable results.

C0304: **Clustering categorical data using a Bayesian mixture of finite mixtures of latent class analysis models**
Presenter: **Bettina Gruen**, WU Vienna University of Economics and Business, Austria
Co-authors: Gertraud Malsiner-Walli, Sylvia Fruehwirth-Schnatter

A Bayesian approach is proposed for model-based clustering of multivariate categorical data where variables are allowed to be associated within clusters, and the number of clusters is unknown. The approach uses a two-layer finite mixture of mixtures model where the cluster distributions are approximated using latent class analysis models. A careful specification of priors with suitable hyperparameter values is crucial to identify the two-layer structure and obtain a parsimonious cluster solution. The Bayesian estimation is outlined based on Markov chain Monte Carlo sampling with the telescoping sampler, and it describes how to obtain an identified clustering model by resolving the label-switching issue. Empirical demonstrations in a simulation study using artificial data as well as a data set on low back pain indicate the good clustering performance of the proposed approach in case hyperparameters are selected to induce sufficient shrinkage.

C0442: **Predictive inference for ecological problems**
Presenter: **Federico Camerlenghi**, University of Milano-Bicocca, Italy
Co-authors: Lorenzo Ghilotti, Tommaso Rigon

It is emphasized that a fundamental goal of science is prediction rather than the explanation of observed facts. This idea has also been pointed out by another statistician, who wrote, "Science cannot limit itself to theorizing about accomplished facts but must foresee". The Bayesian nonparametric approach offers a natural probabilistic framework to address this fundamental issue through the notion of predictive distributions. The purpose is to consider a population of animals composed of different species with unknown proportions and to address prediction problems in this context. An archetypal problem in the species setting is the estimation of the unseen: Given an initial, observable sample from a population, how many new species will be observed in a future sample from the same population? While Bayesian nonparametric methods traditionally concentrate on abundance data, the scenario of incidence data is considered, where the sampling unit is a plot, and one records the incidence (presence or absence) of a species in the plot. A new Bayesian nonparametric approach is developed, designed for incidence data, providing closed-form expression to address several prediction problems, including the estimation of unseen species and population size. The importance of the findings is showcased in facing biodiversity estimation in a large variety of ecological frameworks.

C1178: **Parametric, nonparametric and repulsive mixture models for ecological data**
Presenter: **Eleni Matechou**, University of Kent, United Kingdom

Ecological surveys often track individuals or species to monitor time-varying processes such as migration patterns and changes in behavioral or life states. Mixture models are a suitable and flexible approach for analyzing such data, and they have been used extensively in the field. Parametric, nonparametric, and repulsive mixture models are discussed for different types of ecological data and present results for case studies on species monitored using standard count data, as well as data collected on individuals using new technologies.

CO127 Room BH (SE) 1.05 TIME SERIES MODELING IN FINANCE AND INSURANCE Chair: Edit Rroji

C0287: **Lee-Carter model: Biases and risk in the estimation of the gender mortality gap**
Presenter: **Giovanna Apicella**, University of Udine, Italy
Co-authors: Emilia Di Lorenzo, Gabriella Piscopo, Marilena Sibillo

Implementing social and economic policies addressing the gender longevity gap should be guided by quantitative analyses about how such a gap is expected to evolve over time. The performance of the Lee-Carter (LC) model is investigated to fit the gender gap ratio (GGR), namely the ratio of male death rates to female ones. The analysis is based on the use of a Cox-Ingersoll-Ross (CIR) process to describe the fitting errors of the LC model. In particular, the long-term mean and the volatility of such a process drive the analysis as they are indicators of the long-term fitting attitude of the LC model along with the overall risk of such a model. Systematic evidence is provided on the LC model performance in estimating GGR trends across ages and on a sample of 25 countries, also by making use of functional cluster analysis.

C0288: **The puzzle of carbon allowance spread**
Presenter: **Michele Azzone**, Politecnico di Milano, Italy
Co-authors: Roberto Baviera, Pietro Manzoni

A growing number of contributions in the literature have identified a puzzle in the European carbon allowance (EUA) market. Specifically, a persistent cost-of-carry spread (C-spread) over the risk-free rate has been observed. This is the first explanation of the anomalous C-spread with the credit spread of the corporates involved in the emission trading scheme. Statistical evidence that the C-spread is cointegrated with both this credit spread and the risk-free interest rate is obtained. This finding has a relevant policy implication: The most effective solution to solve the market anomaly is to include the EUA in the list of European Central Bank-eligible collateral for refinancing operations. This change in the ECB monetary policy operations would greatly benefit the carbon market and the EU green transition.

C0399: **Pricing options with a compound CARMA(p,q)-Hawkes model**
Presenter: **Andrea Perchiazzo**, University of Milan, Italy
Co-authors: Lorenzo Mercuri, Edit Rroji

Recently, a novel self-exciting point process has been introduced in the literature, featuring a continuous-time autoregressive moving average intensity process. Such a model, named CARMA(p,q)-Hawkes, extends the traditional Hawkes process by integrating a CARMA(p,q) framework instead of an Ornstein-Uhlenbeck intensity. As a matter of fact, the proposed model maintains the same level of mathematical tractability as the Hawkes process (e.g., infinitesimal generator, backward and forward Kolmogorov equations, joint characteristic function), but it shows enhanced capability in reproducing complex time-dependent structures evident in several market data. Based on this framework, a compound CARMA(p,q)-Hawkes model is proposed, incorporating a random jump size independent of both the counting and intensity processes, which serves as a key component for a new option pricing model. An analysis is conducted to assess the effectiveness of this pricing model in replicating the volatility surface observed in market option data.

C1151: **Identifying the number of latent factors of stochastic volatility models**
Presenter: **Erindi Allaj**, University of Parma, Italy

A procedure is provided to identify the number of latent factors of stochastic volatility models. The methodology relies on the non-parametric Fourier estimation method introduced by a past study and applies to high-frequency data. Based on the Fourier analysis, the latent volatility process is first estimated, and then the volatilities and covariances of the processes are gradually identified, such as volatility of volatility and leverage. The analysis of the eigenvalues spectrum of the Gram matrix can reveal information about the actual number of factors driving the process at hand. The analysis is corroborated by numerical simulations on single and multi-factor models. Finally, the methodology is applied to intraday prices from the S&P 500 index futures.

CO016 Room BH (SE) 1.06 HITEC: ADVANCES IN FINANCE AND STATISTICS Chair: Davide Lauria

C0172: **Sparsity-constrained estimators for graphical models**
Presenter: **Alessandro Fulci**, University of Trento, Italy

Graphical models provide a versatile framework for representing conditional dependence structures among random variables. New methods are proposed for estimating a sparse and shrunk precision matrix in undirected Gaussian graphical models. New approaches are introduced that, besides incorporating the l1-norm penalty, rely on a constraint on the l0-pseudo-norm. Specifically, the newly introduced estimators include the sparsity-constrained Glasso (SCGlasso), the adaptive graphical Lasso (AGlasso), and its sparsity-constrained variant (SCAGlasso). An essential feature of our proposed algorithms is their ability to circumvent the highly non-convex nature associated with the l0-constraint. A comprehensive comparison with established methods like Glasso and the Atan penalty is conducted through simulations, revealing the effectiveness of the l0-

constraint, particularly in model selection and for small sample sizes. Additionally, a real-world application is explored in the context of gene expression, supporting the validity of the proposed approaches.

C0880: **Evaluating the impact of methodological choices on ESG scores**
Presenter: **Sandra Paterlini**, University of Trento, Italy
Co-authors: Matteo Benuzzi, Ozge Sahin

Environmental, social, and governance (ESG) providers aggregate various types of information, particularly key performance indicators (KPIs), to generate comprehensive scores. This aggregation process typically involves numerous binary and continuous variables, often plagued by significant amounts of missing data. The aggregation methodologies employed by these providers can significantly influence the dependence structures within the data. Using data from Refinitiv, a leading ESG provider, insights are provided into the impact of aggregation methodology on the dependence structure between raw and normalized KPIs in the presence of missing data, as well as on the number of KPIs needed to construct representative ESG scores.

C1477: **Distributionally robust optimal portfolios and ESG ambiguity**
Presenter: **Davide Lauria**, University of Bergamo, Italy
Co-authors: Rosella Giacometti, Gabriele Torri

ESG optimal portfolios are investment portfolios that aim to balance financial returns with environmental, social, and governance (ESG) factors. The goal is to construct a portfolio that maximizes returns while incorporating ESG criteria to ensure environmental sustainability, ethical practices, and good governance. However, ESG scores are produced by various rating agencies, each using a unique and proprietary methodology, leading to partially inconsistent evaluations of firms across agencies. This ambiguity, in turn, affects the selection of an ESG optimal portfolio, whose composition can vary significantly depending on the chosen rating agency. The scope is twofold. First, the sensitivity of mean-CVaR and mean-VaR optimal portfolios to the choice of the rating agency is assessed. Then, an ESG index and a distributionally robust ESG portfolio optimization approach are introduced to create optimal portfolios that remain robust regardless of the rating agency's methodology or market conditions.

C1370: **Start-to-low drawdown as a risk measure and its application to levered investor portfolio optimization**
Presenter: **Philipp Staehli**, University of Basel, Switzerland
Co-authors: Dietmar Maringer

Drawdown is an important risk measure in theory and practice. Most drawdown measures use the running peak as the reference point from which to calculate the drawdown. In the first part, the start-to-low drawdown (SLD), referencing the period start instead, is proposed as a relevant measure for levered investors. The characteristics of the SLD are compared with other popular risk measures. In the second part, an application to a levered investor who is additionally subject to regulatory capital requirements as known in the banking or insurance industry is proposed. Such an investor is subject to regulatory sanctions as soon as their own funds no longer cover capital requirements, i.e. even before equity is exhausted. A portfolio optimization objective is developed that considers return, cost of capital and cost of drawdown together: the solvency cost-adjusted return including the cost of drawdown (SCARD). The empirical analysis employs the European insurance industry as an example. The Solvency II standard model is used to calculate capital requirements, and the model of life and non-life insurance companies is constructed from EIOPA market overview data. The characteristics and performance of SCARD-optimal portfolios of the model companies are compared with those following objectives that do not take drawdown into account.

CO077 **Room BH (S) 2.01** MODELLING RISK AND UNCERTAINTY **Chair: Paulo Rodrigues**

C1479: **Monetary policy and growth-at-risk: The role of institutional quality**
Presenter: **Afonso Souto de Moura**, Banco de Portugal and Nova SBE, Portugal
Co-authors: Lorenz Emter, Ralph Setzer, Nico Zorell

The purpose is to analyze how country-specific institutional quality shapes the impact of monetary policy on downside risks to GDP growth in the euro area. Using identified high-frequency shocks in a growth-at-risk framework, it is shown that monetary policy has a higher impact on downside risks in the short term than in the medium term. However, this result for the euro area average hides significant heterogeneity across countries. In economies with weak institutional quality, medium-term growth risks increase substantially following contractionary monetary policy shocks. In contrast, these risks remain relatively stable in countries with high institutional quality. This suggests that improvements in institutional quality could significantly enhance euro area countries' economic resilience and support the smooth transmission of monetary policy.

C1506: **A new decomposition method using expectiles**
Presenter: **Pedro Raposo**, Catolica Lisbon school of business and economics, Portugal
Co-authors: Pedro Portugal, Paulo Rodrigues

A new expectile panel data method is developed for high-dimensional fixed effects estimation in line with prior research, which allows for a wide range of applications in fields such as labor economics, the economics of education, and inequality. It is also shown how the Gelbach decomposition can be validly implemented in the context of panel expectile regressions. Using a unique Portuguese-linked employer-employee dataset, the approach is used to explore the determinants of the gender wage gap over the period 1995-2021. It is found that (i) the gender wage gap is significantly larger in the upper tail; (ii) the difference is mainly explained both in the left and right tail, around 80% by the individual unobserved heterogeneity and 20% by the firm unobserved heterogeneity. Education does not play a significant role in explaining the wage difference between man and woman.

C1528: **A simple but powerful tail index regression**
Presenter: **Paulo Rodrigues**, Universidade Nova de Lisboa, Portugal
Co-authors: Joao Nicolau

A flexible framework is introduced for the estimation of the conditional tail index of heavy-tailed distributions. This framework's tail index is computed from an auxiliary linear regression model that facilitates estimation and inference based on established econometric methods, such as ordinary least squares (OLS), least absolute deviations, or M-estimation. It is shown theoretically and via simulations that OLS provides interesting results. The Monte Carlo results highlight the adequate finite sample properties of the OLS tail index estimator computed from the proposed new framework and contrast its behavior to that of tail index estimates obtained by maximum likelihood estimation of exponential regression models, which is one of the approaches currently in use in the literature. An empirical analysis of the impact of determinants of the conditional left- and right-tail indexes of commodities' return distributions highlights the empirical relevance of our proposed approach. The novel framework's flexibility allows for extensions and generalizations in various directions, empowering researchers and practitioners to straightforwardly explore a wide range of research questions.

C1564: **Risk and heterogeneity in benefits from vocational versus general secondary education**
Presenter: **Hugo Reis**, Banco de Portugal, Portugal

A dynamic model of individual labor market careers (turnover and search, wage development) is estimated on Portuguese panel data of graduates from vocational and general secondary education. It is found that vocational graduates benefit more from the internal labor market than from the external market. This is better for mature individuals than for young individuals. This hurts as among the mature, vocational has a higher lay-off probability. To the common result that vocational education trades early employment advantages for later disadvantages, a decomposition of employment status is added to its dynamic components. To the literature on wage effects, a breakdown of variances in heterogeneity and risk is added.

CO281 Room BH (S) 2.02 ECONOMETRICS OF GROWTH CONVERGENCE AND ENERGY MARKETS Chair: Joachim Schnurbus

C0668: **A review of Phillips-Sul approach-based club convergence tests**
Presenter: **Mateusz Tomal**, Krakow University of Economics, Poland

The Phillips-Sul approach to testing the club convergence hypothesis has attracted considerable research interest in recent years due to its advantages over alternative methods. The aim is to review theoretical papers that extend the Phillips-Sul approach, empirical studies that apply Phillips-Sul approach-based club convergence tests, as well as the software used to execute these methods. The review revealed that, first, the Phillips-Sul approach has seen modifications regarding the procedure of trend extraction from time series, the log t regression and the algorithm clustering panel units into convergence clubs. Second, the Phillips-Sul approach has been widely used not only in economics and finance but also in ecological, energy and health studies. Finally, guidance is provided for further development of the Phillips-Sul approach. This review is useful for researchers and practitioners investigating convergence and club convergence processes.

C1552: **Convergence clubs in the European Union**
Presenter: **Joachim Schnurbus**, University of Passau, Germany
Co-authors: Harry Haupt, Willi Semmler

The convergence of economic growth is analyzed for regions of the European Union and the Eurozone. Using recent approaches to data-driven identification of countries' club membership, considerable variation is found in club composition across countries and time, and a nonparametric approach is used to analyze and explain this variation.

C1616: **Measuring geopolitical risk on energy inflation: A panel quantile approach**
Presenter: **Cristina Amado**, University of Minho, Portugal
Co-authors: Ignacio Garron Vedia, Helena Veiga

A number of the key factors affecting energy prices are addressed from both the demand and the supply side. These factors include geopolitical risks, oil prices, climate change and GDP growth. In order to investigate the effects of these factors and, in particular, of geopolitical risk on the growth-at-risk of energy inflation, a quantile regression is employed in the panel data framework using country-fixed effects. The panel covers country-level indicators for 16 OECD countries from January 1985 to December 2022. The novelty is that the focus is on the tails of the energy inflation distribution, using a methodology that combines quantile regressions and local projections.

C1483: **Sector risk in the European stock market: Navigating the energy transition era**
Presenter: **Jone Ascorbebeitia**, University of the Basque Country UPV/EHU, Spain
Co-authors: Susan Orbe, Eva Ferreira

The energy transition is reshaping stock market dynamics, influencing stock valuations and market trends, and creating new sectoral risks and opportunities, particularly in Europe, where regulatory policies and green initiatives are key drivers. The dependence structure of sectors is examined in the European stock market over the past decade. In particular, the time-varying multivariate dependence is studied between European MSCI sector indexes and special attention is paid to the energy producer and distributor sectors group (P&D) and energy consumer sectors group, obtaining that, as expected, there are high multivariate dependences in recession periods. CoVaR is used to measure the sector risk. The Energy sector is highlighted as the most systemically damaged sector, showing that the risky quantiles of the P&D sectors have the highest negative impact, generating, on average, a 5.4% return loss on the European market in the post-COVID era. Moreover, the results suggest a common spillover effect of sectors on the European market risk. It is obtained that this effect is to reduce market returns by about 3%, not too far from the individual impact of the sectors, which, on average, is around 4%.

CO079 Room BH (S) 2.03 MODELS WITH LARGE DIMENSIONAL AND FUNCTIONAL VARIABLES Chair: Yoosoon Chang

C1213: **Surfing the cross-sectional density of characteristics for factor timing**
Presenter: **Soohun Kim**, KAIST, Korea, South
Co-authors: Yoosoon Chang, Youngmin Choi, Joon Park

The aim is to propose a method that leverages the entire cross-section of stocks rather than concentrating solely on the top and bottom deciles to achieve optimal factor timing performance. By exploiting the timely information embedded in the cross-sectional density of characteristics at a given point in time, the factor timing strategy significantly outperforms traditional long-short portfolios.

C1225: **Risk and monetary policy in a data-rich model**
Presenter: **Haroon Mumtaz**, Queen Mary University of London, United Kingdom

The purpose is to quantify the role of financial conditions and U.S. monetary policy in shaping risk measures associated with a large set of economic indicators. Specifically, a factor-augmented VAR model is estimated with endogenous stochastic volatility, and U.S. financial and monetary policy shocks are isolated. Substantial heterogeneity is found in how risk evolves over the business cycle across economic indicators and across sectors of the economy. Furthermore, preliminary findings reveal that monetary policy can help reduce downside risks.

C1354: **On the source of seasonality in price changes: The role of seasonality in menu costs**
Presenter: **Nao Sudo**, Bank of Japan, Japan

Seasonality is among the most salient features of price changes, but it is notably less analyzed than the seasonality of quantities and the business cycle component of price changes. To fill this gap, the scanner data of 199 categories of goods in Japan is used to empirically study the seasonality of price changes from 1990 to 2021. It is found that the following four features generally hold for most categories: (1) The frequency of price increases and decreases rises in March and September; (2) Seasonal components of the frequency of price changes are negatively correlated with those of the size of price changes; (3) Seasonal components of the inflation rate track seasonal components of net frequency of price changes; (4) The seasonal pattern of the frequency of price changes is stable relative to that of the size of price changes. The pattern is, however, responsive to changes in the category-level annual inflation rate for the year. A simulation analysis is conducted using a simple state-dependent price model, and seasonal cycles in menu costs are shown to play an essential role in generating seasonality of price changes in the data. The nature of seasonal cycles in menu costs and their implications for macroeconomic dynamics are then discussed.

C1623: **Yield curve control policy in Japan: A functional error-correction model approach**
Presenter: **Mototsugu Shintani**, University of Tokyo, Japan
Co-authors: Yoosoon Chang, Joon Park

During the period of unconventional monetary policy, the Bank of Japan conducted the quantitative and qualitative monetary easing policy (QQE) from April 2013 to March 2024. In September 2016, the Bank of Japan introduced the QQE with yield curve control, which was intended to stimulate the economy by flattening the yield curve. In the analysis, a recently developed functional time series method is employed to study how the shape of the yield curve evolved in Japan in response to the introduction of unconventional monetary policy and evaluate its effectiveness.

CO071 Room BH (S) 2.05 MARKOV SWITCHING PROCESSES AND APPLICATIONS Chair: Maddalena Cavicchioli

C0390: **Time-varying identification of monetary policy shocks**
Presenter: **Tomasz Wozniak**, University of Melbourne, Australia
Co-authors: Annika Camehl

A new Bayesian heteroskedastic Markov-switching structural vector autoregression is proposed with data-driven time-varying identification. The model selects alternative exclusion restrictions within regimes and, as a condition for the search, allows identification to be verified through heteroskedasticity. It is shown that US data support time variation in US monetary policy shock identification. In the sample-dominating first regime, systematic monetary policy follows a Taylor rule extended by the term spread, effectively curbing inflation. The second regime, gaining more persistence after the global financial and COVID crises, is characterized by a money-augmented Taylor rule, providing economic stimulus and featuring the liquidity effect.

C0338: **Multivariate Markov switching BEKK models: Filtering, estimation and data analysis**
Presenter: **Jie Cheng**, Keele University, United Kingdom

The focus is on extending the standard multivariate BEKK model, as detailed in an existing study, by allowing both the unconditional correlation and the parameters to be driven by an unobservable Markov chain. In particular, two estimation algorithms are proposed using extended Kalman filters derived from suitable state space representations of the considered model. Numerical examples make evident the effectiveness of the proposed nonlinear estimations. Finally, real-data applications on some financial returns show empirical evidence that the high volatility persistence and correlation changes of such returns can be well explained by estimating multivariate Markov switching BEEK parameters via the two efficient proposed algorithms.

C1256: **Triple-win performance measurement for sustainable supply chains: A Markov-switching decision trees approach**
Presenter: **Fabio Demaria**, University of Modena and Reggio Emilia, Italy
Co-authors: Maddalena Cavicchioli, Ulpiana Kocollari, Federico Bertacchini

The assessment of sustainability within supply chains has become essential for performance measurement, particularly in identifying the primary objectives that sustainability metrics should address. While existing literature often highlights the importance of transparency regarding social and environmental impacts, the core goals of sustainability management are to objectively assess and mitigate risks while improving performance. A significant challenge in sustainability performance measurement and management research is addressed: identifying trade-offs and achieving triple-win outcomes by analyzing economic, social, and environmental data across the supply chain. Advanced techniques that integrate parametric and non-parametric machine learning tools are employed, specifically Markov-switching decision trees, which combine decision tree methods with time-series modeling. This approach is well-suited for uncovering latent patterns and diverse sustainability performance combinations based on various stakeholder expectations. In fact, the states are capable of identifying evolving patterns in sustainability performance that successfully respond to diverse stakeholder priorities. Findings effectively measure sustainability issues through the triple bottom line pillars and introduce a new strategy for sustainability reporting.

C0317: **Bispectral analysis of Markov switching bilinear models**
Presenter: **Maddalena Cavicchioli**, University of Modena and Reggio Emilia, Italy
Co-authors: Ahmed Ghezal, Imane Zemmouri

The purpose is to derive matrix expressions in closed form for the spectral and bispectral densities of Markov switching bilinear models. Under suitable assumptions, the sample estimators of the spectral and bispectral density matrices are proven to be consistent and asymptotically normally distributed. Simulations and empirical applications confirm the validity of the asymptotic properties and the suitability of the proposed methods for the analysis of time series in the frequency domain.

CO389 Room BH (SE) 2.01 PERSPECTIVE IN EXPLORING DEPENDENCY Chair: ShengLi Tzeng

C0801: **Measuring multivariate regression association via spatial signs**
Presenter: **Jia-Han Shih**, National Sun Yat-sen University, Taiwan
Co-authors: Yi-Hau Chen

A regression association measure is proposed, aiming at the predictability of a multivariate outcome $\mathbf{Y} = (Y_1,...,Y_d)$ from a multivariate covariate $\mathbf{X} = (X_1,...,X_p)$. Motivated by existing measures, Kendall's tau is first generalized to measure the association between two random vectors. The generalized Kendall's tau of two independent replications, $\mathbf{Y}$ and $\mathbf{Y}'$, is then used from the conditional distribution of $\mathbf{Y}$ given $\mathbf{X}$, to measure the predictability of $\mathbf{Y}$ from $\mathbf{X}$. The proposed regression association measure can be expressed as the proportion of the variance of some function of $\mathbf{Y}$ that can be explained by $\mathbf{X}$, indicating that the measure has a direct interpretation in terms of predictability. Based on the proposed measure, a conditional regression association measure is further defined, which can be utilized to perform variable selection. Since the measure is constructed based on two independent replications from the conditional distribution, a simple nonparametric estimation method based on the nearest neighbor is available. Simulations are carried out to examine the performance of the proposed variable selection algorithm, and real data examples are analyzed for illustration.

C0837: **A segmentation method for exploring multivariate spatiotemporal data**
Presenter: **ShengLi Tzeng**, National Chung Hsing University, Taiwan

Understanding important structures within complicated multivariate spatiotemporal data presents many challenges. Common approaches treat spatial locations or time points as variables, transforming the problem into multiple time series or collections of maps. However, these transformations overlook correlations between nearby data in space and time, which tend to be more strongly correlated than more distant data pairs. Direct and cross-variogram functions can describe correlations at varying spatiotemporal distances, but assuming stationarity over space or time may be too restrictive. Building on variograms, clustering methods are integrated, including Delaunay triangulation, adjacency-based spatial continuity constraints, spectral clustering, and the k-modes algorithm. The integration segments space into connected regions and time into continuous intervals in a data-driven way, which allows for more effective capture of spatiotemporal relationships. Anomalies that may not align with overall patterns can also be easily identified.

C0838: **An algorithm for estimating threshold boundary regression models**
Presenter: **ChihHao Chang**, National Chengchi University, Taiwan

The threshold boundary regression (TBR) model is introduced for the analysis of datasets with binary or continuous responses. By integrating regression models with threshold boundary functions using explanatory variables, the TBR model constructs linear or nonlinear classifiers, partitioning the responses into two groups, with separate regression models fitted to each group. An ordered iterative algorithm called the TBR-WSVM algorithm is proposed to estimate the TBR model. This algorithm combines weighted support vector machine (WSVM) techniques with maximum likelihood and least-squares methods. Through simulation studies and empirical analyses, the performance of the TBR-WSVM algorithm is assessed. The results indicate that the TBR-WSVM algorithm offers robust estimation and prediction capabilities for linear and nonlinear threshold boundary models.

C0857: **Extreme value analysis using semiparametric spatial zero-inflated models**
Presenter: **Chun-Shu Chen**, National Central University, Taiwan
Co-authors: Chung-Wei Shen, Bu-Ren Hsu

Spatial two-component mixture models are effective for analyzing spatially correlated data with zero inflation. To avoid biases from assuming a specific distribution for response variables, a semiparametric spatial zero-inflated model is utilized. This model presents significant computational challenges, especially with large datasets, due to the high dimensionality of latent spatial variables, complex matrix operations, and slow estimation convergence. To address these issues, a projection-based method is introduced that reduces dimensionality by projecting latent spatial variables onto a lower-dimensional space using selected basis functions. An efficient iterative algorithm is developed for parameter estimation within a generalized estimating equation framework. The optimal number of basis functions is determined via Akaike's information criterion, and the stability of parameter estimates is assessed using the block jackknife method. This approach is validated through simulation studies and applied to Taiwan's 2016 daily rainfall data, demonstrating its practical effectiveness.

CO306 Room BH (SE) 2.05 MACROECONOMIC CONSEQUENCES OF CLIMATE CHANGE Chair: Francesco Simone Lucidi

C0576: **Climate growth-at-risk**
Presenter: **Damiano Di Francesco**, SantÁnna School of Advanced Studies, Italy
Co-authors: Francesco Lamperti, Christian Brownlees

A comprehensive analysis of how climate shocks dynamically influence the entire forecast distribution of GDP growth is offered. This approach is called climate growth-at-risk (Climate GaR). While recent literature on the macroeconomic effects of climate change predominantly focuses on the average trajectory of GDP growth, the dynamic tail risks triggered by climate change have not been adequately addressed. Climate GaR seeks to bridge this gap by allowing for asymmetric impacts across different regions of the GDP growth distribution. Panel quantile local projections are employed on a large panel of countries from 1960 to 2019, finding suggestive evidence that climate shocks negatively affect downside risks to real activity outcomes, especially in the medium term. In analyzing the impacts of climate shocks on GaR across different countries, heterogeneity is revealed in responses between wealthier and poorer nations, as well as between countries with low and high agricultural intensity. By doing so, novel insights are provided into how climate change poses asymmetric risks to the economy, emphasizing the need for policy frameworks to address and mitigate these heightened vulnerabilities.

C1233: **Green assets, risk preferences and portfolio selection with sustainability**
Presenter: **Ibrahim Tahri**, International Institute for Applied Systems Analysis (IIASA), Austria
Co-authors: Lebogang Mateane, Willi Semmler

A streamlined framework specifically designed for risk-averse green investors is presented with the goal of fostering sustainable and environmentally responsible production processes. This framework is crafted to resonate with the increasing global appetite for green investments among a diverse group of market participants, including institutional investors, asset managers, central banks, and sovereign wealth funds. The approach involves profiling green investors who closely monitor and respond to short-term global financial uncertainty, as quantified by the Chicago Board Options Exchange Volatility Index (VIX). The uncertainty is modeled using a two-state Markov process, differentiating between periods of high and low volatility. These distinct states of financial uncertainty drive changes in investors' preferences, shifting between approximated CRRA (Constant Relative Risk Aversion) and IRRA (Increasing Relative Risk Aversion) expected utilities. This shift influences portfolio reallocations, moving investments away from carbon-intensive (brown) assets toward sustainable (green) assets. CRRA expected utility is associated with periods of low financial uncertainty, while IRRA expected utility aligns with high volatility periods. The model incorporates higher moments of green asset returns into the approximated CRRA and IRRA expected utilities, highlighting investors' preference for positively skewed portfolios, particularly those that emphasize renewable energy assets.

C1544: **Global food price, weather shocks, and inventory**
Presenter: **Marta Maria Pisa**, Ghent University, Belgium

The aim is to study how weather affects global food prices, taking into account the role of inventory. Specifically, the purpose is to examine the asymmetries that arise when adverse weather occurs in conjunction with low inventory levels, compared to high inventory levels.

C1690: **Errors in temperature forecasts and energy prices**
Presenter: **Francesco Simone Lucidi**, Federico II University, Italy

The aim is to analyze the impact of temperature forecast errors on European energy prices, with a focus on the natural gas market. Using ensemble temperature forecasts, the study constructs measures of temperature surprises and updates to assess their influence on gas demand and prices. The findings aim to offer insights into how anticipated and unanticipated temperature changes impact the energy sector and the broader European economy.

CO004 Room BH (SE) 2.09 MACRO-FINANCE APPLICATIONS OF QUANTILE VAR MODELS Chair: Simone Manganelli

C0214: **The global financial cycle and macroeconomic tail risks**
Presenter: **Yves Schueler**, Deutsche Bundesbank, Germany

The purpose is to study the link between the global financial cycle and macroeconomic tail risks using quantile vector autoregressions. Contractionary shocks to financial conditions and monetary policy in the United States cause elevated downside risks to growth around the world. By tightening financial conditions globally, these shocks affect the left tail of the conditional output growth distribution more strongly than the center of the distribution. This effect is particularly pronounced for countries with less flexible exchange rate arrangements, higher foreign currency exposures, and higher levels of private sector leverage, suggesting that exchange rate policies and macroprudential policies can mitigate downside risks to growth.

C0553: **Quantifying financial stability trade-offs for monetary policy**
Presenter: **Frederik Lund-Thomsen**, European Central Bank, Germany
Co-authors: Sulkhan Chavleishvili, Manfred Kremer

An empirical approach is proposed to the integration of short- to medium-term financial stability considerations into monetary policy aimed toward achieving conventional macroeconomic stability. A nonlinear vector autoregression is estimated for the euro area covering the real economy, monetary policy, and measures of ex-ante and ex-post systemic risk. Policy implications are derived from scenario analyses, where assumptions about the future paths of the two systemic risk measures imply distinct financial stability trade-offs for monetary policy. Different monetary policy responses are evaluated within an intertemporal cost-benefit analysis that supports asymmetric central bank loss functions and hence, risk management approaches.

C0781: **Estimating multivariate macroeconomic risk**
Presenter: **Maximilian Schroder**, European Central Bank, Germany

Analyzing multivariate macro-financial risk is important to understand how economic shocks affect the probability of tradeoffs. This is particularly relevant for economic agents pursuing multiple objectives. The necessary tools are introduced to study the complex nature of joint macro-financial uncertainty and how structural economic shocks play a key role in shaping the risk outlook across multiple macro-financial aggregates. In addition, inspired by the finance literature, the framework allows summarizing the multivariate risk into simple risk measures that remain consistent with economic agents' preferences. The findings suggest that the effects of shocks are strongly heterogeneous depending on the economic environment and generally stronger in unstable times. In addition, if a shock drives variables of interest in opposite directions, then the effect of the economic shock on the overall balance of risk becomes ambiguous and strongly time-varying. In the case of monetary policy, if the policy is aimed at mitigating risks to the outlook, timing the policy response appropriately is key. A wait-and-see approach during the early recovery post-pandemic, coupled with more decisive action as the economy stabilized, is supported by the empirical results.

C1657: **The effects of monetary policy on macroeconomic risk**
Presenter: **Nicolo Maffei Faccioli**, Norges Bank, Norway

Monetary policy expansions significantly reduce macroeconomic downside risk, measured as the difference between the median and the 5th percentile of the industrial production growth forecast distribution. However, the effects are smaller in magnitude than those of credit spread shocks, which is found to be a major driver of fluctuations in downside risk. Consequently, large policy interventions are required to stabilize risk originating from the financial sector, with undesirable consequences in terms of both price and output stability. These findings are obtained using US data and a novel econometric approach which combines quantile regressions and Structural VAR analysis.

CO356 **Room BH (SE) 2.12** **TOPICS IN ECONOMIC POLICY AND INTERNATIONAL FINANCE** **Chair: Christian Proano**

C1453: **How fitting is one-size-fits-all: Revisiting the dynamic effects of ECB's interest policy on Euro area countries**
Presenter: **Maybrit Waechter**, University of Bamberg, Germany

The purpose is to revisit the one-size-fits-all challenge posed by the European Central Banks (ECB) monetary policy within the heterogeneous economic landscape of the Eurozone. Using a dataset spanning from 1991Q1 to 2022Q1 across EU-11 countries, country-specific hypothetical Taylor rates are computed, and the dynamic effects of the Taylor Rate Gap (TRGAP) - the difference between these rates and the actual ECB policy rate - are examined on GDP growth, inflation, unemployment, and government debt. Employing local projections and panel OLS regressions, findings reveal that the TRGAP negatively impacts economic growth, with this effect being more pronounced in periphery countries compared to core countries. The analysis highlights the limitations of a uniform monetary policy in addressing the diverse economic conditions within the Eurozone, suggesting the need for a more tailored approach to foster balanced and sustainable growth across the region.

C1466: **Output gap uncertainty, fiscal policy and risk premia under endogenous credibility**
Presenter: **Christian Proano**, University of Bamberg, Germany

The implications of output gap uncertainty (resulting from the "unobservability" of potential output and the use of "imperfect" techniques such as the HP filter for its approximation/estimation) are investigated for the conduct of fiscal policy using a small-scale macroeconomic model with boundedly rational agents along the lines of another study. More specifically, agents are assumed to be unable to know or estimate accurately the true potential output level given their bounded rationality. Instead, they try to approximate it by detrending the actual, observable output using an adaptive updating mechanism or the proper Hodrick and Prescott filter. As it is well known, these estimates will suffer, by construction, from an end-point bias that may lead to a systematic underestimation of the true difference between the actual and the potential output level, i.e. of the output gap, and by extension, to an unintended procyclicality in the conduct of fiscal policy. This, in turn, will affect the government's credibility - endogenized through a binary choice approach along the lines of another study and, by extension, the risk premium on government bonds.

C1540: **Commodity price shocks, geopolitical risk, and macroeconomic activity**
Presenter: **Leonardo Quero Virla**, Otto-Friedrich-Universitaet Bamberg, Germany
Co-authors: Christian Proano

The interaction between fluctuations in commodity prices and macroeconomic activity has been a recurrent topic in empirical macroeconomics for at least four decades. Such interaction is often linked to geopolitical events, which result in large price and quantity movements in international commodity markets. The aim is to contribute to the literature by exploring the joint dynamics of commodity prices, geopolitical risk and economic activity at the global and Euro Area levels. In the frequency domain, through the Breitung-Candelon test, limited causality is found between commodity prices and global economic activity (in both directions) at certain frequencies, but the geopolitical risk was not relevant in any case. In the time domain, the recursive SVAR and local projections results show that while geopolitical risk produces a general decrease in economic activity, commodity price shocks are contractionary for developed economies and expansionary for emerging ones. However, regime-based local projections show that the level of geopolitical risk influences the size and sign of responses to geopolitical risk shocks. Additionally, a Bayesian SVAR analysis with sign and zero restrictions and hierarchical prior selection suggests that both geopolitical risk and commodity prices have an important role in explaining fluctuations in the level of output and economic expectations in the Euro Area.

C1556: **Borrower-and lender-based macroprudential policies: How do they affect the transmission mechanism of fiscal policy**
Presenter: **Lebogang Mateane**, University of Cape Town, South Africa
Co-authors: Christian Proano

The aim is to investigate how borrower and lender-based macroprudential policies affect the transmission mechanism of fiscal policy. A dynamic stochastic general equilibrium model is constructed with a fiscal and banking sector. It is found that upon impact, a government spending shock generates an increase in aggregate consumption; however, there is no permanent increase in aggregate consumption. This is within a setting where entrepreneurs are borrowers; they exhibit an increase in consumption upon impact, whereas patient households are savers and they exhibit a decrease in consumption. The findings are aligned with models that incorporate household heterogeneity. Specifically, non-Ricardian and Ricardian households in the class of two agent new Keynesian models. It is also found that if policy authorities increase the stringency of borrower-based macroprudential policy and, in response to a government spending shock, they stabilize pro-cyclical bank lending. Results are robust to the lender-based macroprudential policy. Particularly, to a higher degree of responsiveness of the time-varying loan-to-value ratio of entrepreneurs to the capital-to-debt exposure of banks. In an environment of technology shocks and with policy authorities increasing the stringency of borrower-based macroprudential policy, we find that they stabilize pro-cyclical bank lending.

CC442 **Room BH (SE) 1.01** **BAYESIAN ECONOMETRICS** **Chair: Rodney Strachan**

C0289: **Inflation dynamics during the COVID-19 era: A high-frequency approach**
Presenter: **Simon Smith**, Federal Reserve Board, United States
Co-authors: Hie Joo Ahn

Exploiting a novel weekly panel dataset of prices and quantities across 14 sectors and the U.S. aggregate since January 2020, high-frequency inflation dynamics are studied, with a particular focus on the Phillips curve and the pass-through of monetary policy. The novel dataset holds significant out-of-sample predictability for official inflation and retail sales measures, suggesting its usefulness for the inference of aggregate dynamics. The model of Bayesian break detection suggests the Phillips curve steepened from October 2020 to July 2021 and then flattened afterwards. These shifts are detected in real-time, providing timely signals about changes in inflation trends and uncertainty. Last, monetary policy rapidly passed through to price and quantity changes during the recovery after the pandemic recession, likely preventing a deflationary spiral.

C0441: **Bayesian dynamic graphical models for large vector autoregressions with time-varying parameters and volatility**
Presenter: **Seyma Vahap**, Kings College London, United Kingdom

A proposed Bayesian dynamic graphical modelling (BDGM) approach has an important property of splitting a complex and high-dimensional vector autoregressive model with time-varying parameters and volatility discounting into locally structured sparse components. This high-dimensional model has an extensive collection of predictors, most of which make a small contribution to the overall power of the model. Still, it is unclear which predictors are relatively more important. A key role for local model specification and computations is the idea of pairwise conditional independence structure. This approach is achieved by developing an efficient Bayesian graphical variable selection method that can be applied recursively in parallel using a Gray code algorithm. The BDGM approach is applied to ten quarterly U.S. macroeconomic and financial variables. The results of posterior model probabilities over the space of all competing models show that there is considerable model uncertainty within the best-selected models. Then, a Bayesian model averaging (BMA) is performed to forecast the variables of interest over a pseudo out-of-sample forecast period 1984:Q2-2022:Q3. Comparing out-of-sample forecast performances shows that the joint model with BMA outperforms the joint model with the highest posterior probability over the majority of forecast horizons.

C1401: **Flexible Bayesian quantile analysis of residential rental rates**
Presenter: **Mohammad Arshad Rahman**, Indian Institute of Technology Kanpur, India
Co-authors: Shubham Karnawat, Ivan Jeliazkov, Angela Vossmeyer

A random effects quantile regression model is developed for panel data that allows for increased distributional flexibility, multivariate heterogeneity, and time-invariant covariates in situations where mean regression may be unsuitable. The approach is Bayesian and builds upon the generalized asymmetric Laplace distribution to decouple the modeling of skewness from the quantile parameter. An efficient simulation-based estimation algorithm is derived; its properties and performance are demonstrated in targeted simulation studies and employed in the computation of marginal likelihoods to enable formal Bayesian model comparisons. The methodology is applied in a study of U.S. residential rental rates following the Global Financial Crisis. Empirical results provide interesting insights on the interaction between rents and economic, demographic and policy variables, weigh in on key modeling features, and overwhelmingly support the additional flexibility at nearly all quantiles and across several subsamples. The practical differences that arise as a result of allowing for flexible modeling can be nontrivial, especially for quantiles away from the median.

C1333: **Singular vector autoregressions**
Presenter: **Rodney Strachan**, The University of Queensland, Australia
Co-authors: Eric Eisenstat

Methods are developed for the empirical analysis of singular processes. A strong rationale and a well-developed theoretical framework and as it is shown that empirical support exists for multivariate time series with a singular spectral density. A singular spectral density is consistent with the economic theory underlying, for example, DSGE models, in which the number of variables is greater than the number of structural shocks. This assumption guarantees the existence of a finite order VAR representation, but a unique probability density function does not exist with respect to the Lebesgue measure. A density on a compact submanifold is therefore defined with respect to the Hausdorff measure and, in a Bayesian framework, an HMC algorithm is developed that jointly samples coefficients, lag length, and the number of shocks. The proposed framework is used to carry out a structural analysis of the US macroeconomy with COVID-19 shocks.

CC496 Room BH (SE) 2.10 FINANCIAL ECONOMETRICS I Chair: Toshiaki Watanabe

C1653: **A GARCH model with two volatility components and two stochastic factors**
Presenter: **Luca Vincenzo Ballestra**, Alma Mater Studiorum University of Bologna, Italy
Co-authors: Enzo DInnocenzo, Christian Tezza

Several studies provide support for stochastic volatility specifications that incorporate two sources of uncertainty, emphasizing that a simple one-factor model inadequately captures the dynamic dependencies in financial data. A GARCH model with two distinct components is proposed, each driven by an independent, unobservable innovation factor. By introducing a second innovation factor, the model's ability to represent market dynamics is enhanced, resulting in a more accurate description of the variance process and the implied volatility surface. The model improves the two-factor GARCH model recently proposed by allowing spillovers between the volatility components rather than treating them as independent, and sufficient conditions are established for stationarity and ergodicity. Results demonstrate that the approach outperforms existing models in capturing S&P 500 index returns, both in-sample and out-of-sample. Additionally, it shows comparable performance to the prior study's model when pricing S&P500 options.

C1250: **A safe haven index**
Presenter: **Thomas Dimpfl**, University of Hohenheim, Germany
Co-authors: Dirk Baur, Javier Pena

The literature has identified several safe haven assets, but largely overlooked their collective dynamics. A novel safe haven index (SHI) designed to track the dynamics of the safe haven asset market is introduced. The index is based on the first principal component of all constituent assets and protects against extreme stock market shocks. Applying a linear factor pricing model, it is demonstrated that the SHI is a priced risk factor and exhibits characteristics akin to the HML factor in Fama and French's model.

C1440: **A semiparametric semi-strong FARIMA with heteroskedastic errors applied to energy market**
Presenter: **Aliyu Abubakar Musa**, University of Paderborn, Germany, Germany
Co-authors: Yuanhua Feng

Usually, the FARIMA and the SEMIFAR (semiparametric fractional autoregression) models are assumed to be linear with i.i.d. errors. The FARIMA-GARCH and SEMIFAR-GARCH models provide semi-strong, non-linear extensions of them with uncorrelated GARCH errors. A SEMIFAR-SemiGARCH model is proposed by introducing a latent smooth scale function into the uncorrelated GARCH errors. This provides an approach with trend and long-memory in the mean part, as well as a (latent) smooth scale function and stationary volatility in the errors, where the last component can be analyzed by different GARCH models. The properties of this novel model and its estimation theory are investigated. A multi-step algorithm that combines the SEMIFAR and SemiGARCH estimation procedures has been developed for practical implementation. The performance of this algorithm is justified. This proposal is applied to a selected time series in the energy market. Results show that the proposed model is useful in practice and can be applied to forecast different stationary and non-stationary components in a time series. It can also be extended to include long memory in volatility.

C0368: **Matrix-based prediction approach for intraday instantaneous volatility vector**
Presenter: **Sung Hoon Choi**, University of Connecticut, United States
Co-authors: Donggyu Kim

A novel method is introduced for predicting intraday instantaneous volatility based on Ito semimartingale models using high-frequency financial data. Several studies have highlighted stylized volatility time series features, such as interday auto-regressive dynamics and the intraday U-shaped pattern. To accommodate these volatility features, an interday-by-intraday instantaneous volatility matrix process is proposed that can be decomposed into low-rank conditional expected instantaneous volatility and noise matrices. To predict the low-rank conditional expected instantaneous volatility matrix, the two-side projected-PCA (TIP-PCA) procedure is proposed. Asymptotic properties of the proposed estimators are established, and a simulation study is conducted to assess the finite sample performance of the proposed prediction method. Finally, the TIP-PCA method is applied to an out-of-sample instantaneous volatility vector prediction study using high-frequency data from the S&P 500 index and 11 sector index funds.

Sunday 15.12.2024 10:55 - 12:10 Parallel Session G – CFECMStatistics2024

CO085 Room S-1.01 HITEC: TOPICS IN FINANCIAL ECONOMETRICS AND REGIME SWITCHING MODELS Chair: Leopold Soegner

C0864: **Monitoring structural breaks in vector autoregressive models**
Presenter: **Masoud Abdollahi**, Institute for Advanced Studies, Austria
Co-authors: Leopold Soegner

A closed-end monitoring method is developed to perform online breakpoint detection in the non-stationary case. Specifically, an error correction model is considered where structural breaks may occur in cointegrating and/or adjustment vectors, where the cointegration rank either remains the same or changes. Lagrange-multiplier tests are developed to monitor these structural breaks. Following a calibration period used to estimate the model parameters, monitoring starts, and it halts the first time the test statistic exceeds the corresponding critical value. Through an extensive simulation study, the performance of the monitoring procedure is thoroughly investigated.

C1044: **A change point test for a gradual change in the Poisson INARCH(1)-process**
Presenter: **Florian Schirra**, Fraunhofer ITWM, Germany
Co-authors: Stefanie Schwaar, Joern Sass

Change point detection methods are a common tool to identify structural changes in the distribution of time series. In recent years, there has been progress in detecting changes within times series in countable spaces, e.g. the natural numbers. For a number of applications, such as outbreak detection of infectious diseases, modeling a gradual change could be valuable. Such count time series can be modeled by Poisson INARCH(1) processes. One possibility to model gradual changes is by introducing a non-linear time-dependent factor in the intensity function of a Poisson INARCH(1) process. This additional factor characterizes the gradual change after the change point. The distribution of a test statistic based on partial sums of weighted residuals still has a limiting distribution given by the Gumbel extreme value distribution under the null hypothesis. Under the alternative, consistency holds for certain assumptions.

C1292: **Regime switching for dynamic equicorrelation**
Presenter: **Yannick Le Pen**, Universite Paris Dauphine, France
Co-authors: Arthur Thomas, Zakaria Moussa

The purpose is to introduce a regime-switching dynamic, inspired by Pelletier's Markov switching conditional correlation model, into Engle and Kelly's dynamic equicorrelation (DECO) model. The DECO model is suitable for a vast array of correlations, segmenting the correlation matrix into blocks with equal conditional correlations, while inter-block correlations can be unique or block-specific. Pelletier's regime-switching model, which builds on Engle's model, is limited to a smaller set of returns. The regime-switching dynamic equicorrelation (RSDECO) model is thus well-suited for modeling large conditional correlations with regime shifts. Through extensive simulations, RSDECO is demonstrated to accurately capture the true correlation levels across numerous variables. The model is exemplified by estimating daily correlations among commodity, stock, and bond returns from April 1, 2000, to December 29, 2022. Significant shifts are identified in correlation patterns, notably around September 15, 2008, and substantial alterations are observed in the correlations among the three asset classes post-2020.

CO339 Room S-1.04 APPLICATIONS OF POINT PROCESS AND RELATED MODELS FOR COUNTS (VIRTUAL) Chair: Satish Iyengar

C0347: **Distributional theory for the Conway-Maxwell-Poisson distribution**
Presenter: **Robert Gaunt**, The University of Manchester, United Kingdom

The Conway-Maxwell-Poisson (CMP) distribution is a two-parameter generalization of the Poisson distribution that can be used to model data that are under- or over-dispersed relative to the Poisson distribution. It is particularly widely used in modelling count data. A review of some of the most important distributional properties of the CMP distribution is provided, with a particular focus on the matter of approximating the normalizing constant.

C0675: **Relation between distribution of number of photons detected and interpolation function of Bernoulli-Barnes polynomials**
Presenter: **Burcin Simsek**, Bristol-Myers Squibb, United States

Using the binomial theorem, some novel identities are provided in the same type as those provided in a prior study of photon counts in two-photon laser scanning microscopy. By using this equation and other identities, some new formulas and relations, including the Barnes zeta functions and the Bernoulli-Barnes polynomials and numbers, are derived. A Raabe-type formula is also proven related to these polynomials and numbers. Moreover, applications of these identities and formulas are investigated to stochastic point processes.

C0683: **Early indicators of degradation of materials with applications to batteries**
Presenter: **Satish Iyengar**, University of Pittsburgh, United States

The degradation of materials is a common phenomenon that can lead to failure and require considerable repair. Rechargeable battery-powered devices have become very common, hence the interest in better understanding degradation. These studies are complicated by nonlinear features of degradation patterns. The use of diffusion processes are studied to approximate recent proposals involving compound Poisson process inputs to model transient phenomena.

CO401 Room S-1.27 MACHINE LEARNING AND GLOBAL HEALTH Chair: Alexandra Blenkinsop

C0249: **Detecting and leveraging node-level information in network inference**
Presenter: **Xiaoyue Xi**, Univerisity of Cambridge, United Kingdom
Co-authors: Helene Ruffieux

Bayesian graphical models are powerful tools to infer complex relationships in high dimensions yet are often fraught with computational and statistical challenges. If exploited in a principled way, the increasing information collected alongside the data of primary interest constitutes an opportunity to mitigate these difficulties by guiding the detection of dependence structures. For instance, gene network inference may be informed using publicly available summary statistics on the regulation of genes by genetic variants. The aim is to present a novel Gaussian graphical modelling framework to identify and leverage information on the centrality of nodes in conditional independence graphs. Specifically, a fully joint hierarchical model is considered to simultaneously infer (1) sparse precision matrices and (2) the relevance of node-level information for uncovering the sought-after network structure. Such information is encoded as candidate auxiliary variables using a spike-and-slab sub-model on the propensity of nodes to be hubs, which allows hypothesis-free selection and interpretation of a sparse subset of relevant variables. A variational expectation conditional maximization algorithm is developed that scales inference to hundreds of samples, nodes and auxiliary variables. The advantages of the approach are illustrated and exploited in simulations and in a gene network study, which identifies hub genes involved in biological pathways relevant to immune-mediated diseases.

C0307: **Fast heavy-tail count models for automated probabilistic computing and pandemic preparedness**
Presenter: **Oliver Ratmann**, Imperial College, United Kingdom

The aim is to present new heavy-tail models for discrete count data, including the Zipf distribution and the beta-negative-binomial distribution (BNB) that we contributed for fast inference to Stans C++ Math library, also enabling GPU library support. The Zipf and BNB distributions are, respectively, one and three-parameter distributions to flexibly model scale-free and/or heavy-tail behaviour of a large class of outcomes in generalised linear regression frameworks. These models are implemented and applied to infer how the structure of human contact networks evolved during the COVID-19 pandemic using one of the largest ever, individual-level longitudinal contact survey studies comprising n 50,000 participants in Germany from early 2020 to the end of 2021. The C++ implementations in Stans Math library allow tracking individual-level heterogeneity in contact behavior by age, gender, sub-national geographical areas and attitudes over four pandemic phases. It characterizes which sub-populations high-degree hubs re-emerged fastest and how their relative contribution to all contacts changed over time. At the network level, it is shown how time trends in compounding individual-level behavior are associated with phase transitions in how effectively pandemic spread can be disrupted by estimating population-level scale-free network properties and predicting the existence of giant network components.

C0671: **Sequential decision-making in public health**
Presenter: **Mengyan Zhang**, University of Oxford, United Kingdom

Sequential decision-making methods in machine learning have vital applications in public health. By adaptively making decisions, collecting samples, and learning models, the environment can be understood more efficiently with fewer data points. For instance, policymakers can allocate limited testing resources to maximize information about disease distribution. This decision-making process is modeled as an iterative node classification problem on an undirected, unweighted graph, where nodes represent locations and edges indicate the movement of infectious agents. Findings can aid in designing cost-effective surveillance policies for emerging and endemic pathogens, accelerating disease detection in resource-constrained settings. Moreover, the causal structure is adaptively learnt while optimizing targets through interventions. For example, to minimize HIV viral load by selecting different treatments, graph agnostic causal Bayesian optimization is proposed, an algorithm that actively discovers the causal structure to achieve optimal outcomes. Additionally, methods are developed to address imperfect feedback challenges in public health applications, including non-response bias in survey design and aggregated feedback. The aim is to enhance public health strategies by integrating advanced machine learning techniques, ultimately contributing to more effective and efficient disease control and prevention.

CO367 Room K0.18 NOVEL STATISTICAL TOOLS FOR BIOMEDICAL RESEARCH Chair: Claudia Solis-Lemus

C1418: **Bayesian chain graph model for microbiome data**
Presenter: **Claudia Solis-Lemus**, , United States

A novel Bayesian chain graph model is introduced to infer a sparse network structure with nodes for responses and for predictors with applications to microbiome research. Directed edges between a predictor and a response represent conditional links, and undirected edges among responses represent correlations. Specifically, the model can estimate a microbial network that represents the dependence structure of a multivariate response (e.g. abundances of microbes) while simultaneously estimating the effect of a set of predictors that influence the network (e.g. diet, weather, experimental treatments). In addition, the method produces a sparse interpretable graph via LASSO penalization, which can become adaptive so that different shrinkage can be applied to different edges. Furthermore, the model is able to equally handle small and big data and is computationally inexpensive through an efficient Gibbs sampling algorithm. With hierarchical structure, the model is extended to binary, counting and compositional responses by adding an appropriate sampling distribution to the core Normal model.

C1507: **Bayesian joint models for longitudinal multimorbidity analysis**
Presenter: **Sida Chen**, MRC BSU University of Cambridge, United Kingdom
Co-authors: Danilo Alvares, Chris Jackson, Sylvia Richardson, Jessica Barrett

Multistate models provide a useful framework for modelling complex event history data in clinical settings. They have recently been extended to the joint modelling framework to appropriately handle endogenous longitudinal covariates, such as repeatedly measured biomarkers, which are informative about health status and disease progression. However, the practical application of such joint models faces considerable computational challenges. Motivated by a longitudinal multimorbidity analysis of large-scale UK health records, novel Bayesian inference approaches are introduced for these models that are capable of handling complex multistate processes and large datasets with straightforward implementation. Simulation studies confirm the feasibility of the proposed approaches, with notable advantages in computational efficiency compared to the standard Bayesian estimation strategy. The approaches are used to analyze the coevolution of routinely measured systolic blood pressure (SBP) and the progression of three important chronic conditions based on a large dataset from the clinical practice research datalink aurum database. The analysis reveals distinct and previously lesser-known association structures between SBP and different disease transitions.

C1101: **IsoBayes: A Bayesian approach for single-isoform proteomics inference**
Presenter: **Simone Tiberi**, University of Bologna, Italy
Co-authors: Jordy Bollon

Inferring proteins is a crucial step in biomedical research. At present, proteins are indirectly measured via peptides. However, this process is noisy because most peptides are shared across multiple proteins; furthermore, peptides may also be erroneously detected. As a consequence, studying proteins is challenging, and inferences can be inaccurate. IsoBayes, a novel Bayesian statistical method, is described for protein inference. The goal is to disentangle the biological variability (of interest) from the technical noise arising from the measurement process (nuisance). To this aim, a two-layer latent variable approach is designed where: first, it is sampled if a peptide has been correctly detected, and second, the abundance of such selected peptides is allocated across the protein(s) they are compatible with. This framework enables us, starting from peptide-level data, to recover protein-level information (i.e., presence and abundance). Furthermore, proteomics and transcriptomics data are integrated to enhance the information available. In order to validate the approach, comprehensive benchmarks are designed based on simulated and real datasets, where IsoBayes displays good sensitivity and specificity when detecting proteins and where its estimated abundances highly correlate with the ground truth. Importantly, the method is flexible and works with peptide identifications obtained by any proteomics tool, and it is distributed open-access as a Bioconductor R package.

CO253 Room K0.19 ADVANCE IN MODERN DATA ANALYSIS Chair: Daren Wang

C0769: **Graph-based inference for random effects in high-dimensional linear mixed models**
Presenter: **Lynna Chu**, Iowa State University, United States

Linear random effects model are widely used to analyze correlated or clustered data. A non-parametric approach is proposed to test whether a collection of random effects is zero in models where the fixed effects are high-dimensional. The test statistic is constructed from a similarity graph, such as the minimum spanning tree, and measures the extent to which accounting for correlation improves the ability of a set of fixed effects to predict a response. The asymptotic null distribution of the test statistic is derived and shown to work well under finite samples. The proposed method is assessed using various simulation studies.

C1295: **Parameter inference for partially observed, implicitly defined simulation models**
Presenter: **Joonha Park**, University of Kansas, United States

In many applications, a stochastic system is studied using a model implicitly defined via a simulator. A simulation-based parameter inference method is developed for such implicitly defined models where partial or noisy observations are available. The method differs from traditional likelihood-based inference in that it uses a simulation metamodel for the distribution of a log-likelihood estimator, which is built on a local asymptotic normality (LAN) property. The use of a simulation metamodel enables scalable parameter estimation and uncertainty quantification with increasing data size. The method is demonstrated using numerical examples, including a mechanistic model for the population dynamics of an infectious disease.

C1624: **Generative modeling via hierarchical tensor sketching**
Presenter: **Yifan Peng**, University of Chicago, United States
Co-authors: yian chen, Miles Stoudenmire, Yuehaw Khoo

A hierarchical tensor-network approach is proposed for approximating high-dimensional probability density via empirical distribution. This leverages randomized singular value decomposition (SVD) techniques and involves solving linear equations for tensor cores in this tensor network. The complexity of the resulting algorithm scales linearly in the dimension of the high-dimensional density. An analysis of estimation error demonstrates the effectiveness of this method through several numerical experiments.

CO251 **Room K0.20** RECENT ADVANCES IN SUFFICIENT DIMENSION REDUCTION **Chair: Wei Luo**

C0270: **Fast fitting of Gaussian mixture model via dimension reduction**
Presenter: **Wei Luo**, Zhejiang University, China
Co-authors: Yin Jin

The Gaussian mixture model (GMM) is a widely applied clustering technique. Commonly, GMM is fitted by the maximal likelihood approach that involves non-convex minimization, which becomes computationally challenging, especially for large-dimensional data. A fast two-step approach is proposed to fit GMM with the aid of an intrinsic low-dimensional data structure for clustering under additional constraints on the heterogeneity of GMM. In the first step, a simple moment-based method is proposed to recover the low-dimensional data, given that the rest of the data are normally distributed and thus redundant for clustering. GMM is then fit using the reduced data in the second step, which is computationally more feasible than the original GMM due to the lower dimensionality. Under the sparsity assumption on the clustering pattern, the approach can be generalized under the ultrahigh-dimensional settings. With the aid of appropriate pseudo data, it can also be embedded under a general framework of sufficient dimension reduction, which encompasses more methods to recover the low-dimensional structure of GMM in the future. The numerical studies are presented at the end.

C0532: **Dimension reduction for extreme regression via contour projection**
Presenter: **Jing Zeng**, University of Science and Technology of China, China

In extreme regression problems, a primary objective is to infer the extreme values of the response given a set of predictors. The high dimensionality and heavy-tailedness of predictors limit the applicability of classical tools for inferring conditional extremes. The focus is on the central extreme subspace (CES), whose existence and uniqueness are guaranteed under fairly mild conditions. By projecting the data onto CES, the dimension of the predictors is reduced while all information for inferring conditional extremes is retained, which effectively addresses the high dimensionality issue. The novel COPSE method is proposed to estimate CES by utilizing the contour projection. Notably, COPES is robust against heavy-tailed predictors. The theoretical justification for the consistency of COPES is established. Overall, the proposal not only extends the toolkit for extreme regression but also broadens the scope of the dimension reduction techniques. The effectiveness of the proposal is demonstrated through extensive simulation studies and an application to Chinese stock market data.

C0605: **Dimension reduction for multivariate time series**
Presenter: **Andreas Artemiou**, University of Limassol, Cyprus

The aim is to present two new methods for sufficient dimension reduction in multivariate time series. Following recent advances in the field of sufficient dimension reduction for multivariate time series, it is demonstrated how one can improve the performance of the existing methodologies by using a time series version of the sliced inverse mean difference and by improving the robustness to outliers by using a time series version of the sliced inverse median difference regression.

C1726: **Nonconvex-regularized integrative sufficient dimension reduction for multi-source data**
Presenter: **Shanshan Ding**, University of Delaware, United States

As advances in high-throughput technology significantly expand data availability, integrative analysis of multiple data sources has become an increasingly important tool for biomedical studies. An integrative and nonconvex-regularized sufficient dimension reduction method is proposed to achieve simultaneous dimension reduction and variable selection for multi-source data analysis in high dimensions. The proposed method aims to extract sufficient information in a supervised fashion, and the asymptotic results establish a new theory for integrative sufficient dimension reduction and allow the number of predictors in each data source to increase exponentially fast with sample size. The promising performance of the integrative estimator and efficient numerical algorithms is demonstrated through simulation and real data examples.

CO209 **Room K0.50** ADVANCES IN OPTIMAL EXPERIMENTAL DESIGN **Chair: Sergio Pozuelo Campos**

C0559: **Optimizing active power in electrical systems through optimal experimental design with piezoelectric paints**
Presenter: **Ricardo Negrete Gallego**, University of Castilla-La Mancha, Spain
Co-authors: Irene Garcia-C. Gutierrez, Sergio Pozuelo Campos

Maximizing active power is a fundamental objective in various electrical and energy systems. An innovative methodology is presented based on optimal experimental design (OED) to adjust the experimental setup of an active power model of piezoelectric paints. By using advanced OED techniques, key parameters impacting the generation and distribution of active power are optimized, thereby significantly improving system efficiency. Moreover, some variables related to the used paint quantity have a high cost, so the use of the OED tools allows the collection of most of the information in each experimental unit, reducing thus this cost. A significant advancement is represented in this field, providing a solid foundation for future research and practical applications in active power optimization.

C0681: **Optimal designs for the Baranyi model with two controllable variables**
Presenter: **Alba Munoz**, Universidad de Castilla-La Mancha, Spain
Co-authors: Victor Casero-Alonso, Mariano Amo-Salas

In predictive microbiology, the Ratkowsky square root model is known as a model that describes the influence of suboptimal growth temperatures on the maximum specific growth rate of microorganisms. By integrating the square root model into Baranyi's model, this one becomes a secondary model, dependent on time and temperature. D- and c- optimal experimental designs are provided for the Baranyi model with two controllable variables and considering different perspectives. These designs are compared by means of efficiency with those provided in the literature.

C0919: **Augmented designs to choose between constant absolute and relative errors in regression models**
Presenter: **Samantha Leorato**, University of Milan, Italy
Co-authors: Chiara Tommasi, Carlos de la Calle-Arroyo, Licesio Rodriguez-Aragon

In experimental sciences, such as chemistry, the measurement error may be homoscedastic or heteroscedastic. The data should be collected with the goal of identifying the right error variance structure, as an incorrectly specified model would lead to wrong conclusions. A design criterion that reflects this goal is the KL-optimality. Frequently, however, KL-optimum designs are completely inefficient for other inferential purposes such as precise estimation. In this case, the addition of some experimental points might be convenient. The focus is on the enrichment of a design through the inclusion of some additional support points with the goal of guaranteeing a minimum KL-efficiency to be able to optimally choose between different variance specifications. This strategy is also useful for modifying a design that is already available, for instance, a D-optimal design, to manage the problem of correct error variance specification.

CO204 Room K2.31 (Nash Lec. Theatre) INFERENCE FOR FEDERATED LEARNING AND SYNTHETIC CONTROL Chair: Jia Gu

C1444: **Statistical inference for decentralized federated learning**
Presenter: **Jia Gu**, Center for Data Science, Zhejiang University, China
Co-authors: Songxi Chen

The purpose is to consider decentralized federated learning (FL) under heterogeneous distributions among distributed clients or data blocks for the M-estimation. The mean squared error and consensus error across the estimators from different clients via the decentralized stochastic gradient descent algorithm are derived. The asymptotic normality of the Polyak-Ruppert(PR) averaged estimator in the decentralized distributed setting is attained, which shows that its statistical efficiency comes at a cost as it is more restrictive on the number of clients than that in the distributed M-estimation. To overcome the restriction, a one-step estimator is proposed, which permits a much larger number of clients while still achieving the same efficiency as the original PR-averaged estimator in the non-distributed setting. The confidence regions based on both the PR-averaged estimator and the proposed one-step estimator are constructed to facilitate statistical inference for decentralized federated learning.

C1486: **Dynamic synthetic control method for semiparametric time-varying models**
Presenter: **Shouxia Wang**, Shanghai University of Finance and Economics, China
Co-authors: Songxi Chen, Xiangyu Zheng

Motivated by evaluating the treatment effects of a policy for nonlinear time-varying confounding variables, a dynamic synthetic control (DSC) method is proposed under the semiparametric time-varying additive autoregression outcome model. The proposed method allows for micro-level data with nonlinear time-varying confounders, multiple treated units and spatial correlations in the data. Spline-back-fitted-kernel estimation method is used to obtain good estimations of the unknown additive functions, which are then used for matching when the DSC weights are constructed. The DSC weights are constructed by the empirical likelihood, guaranteeing a unique solution and a consistent estimation of the average treatment effect on the treated group. The semiparametric additive model provides more flexibility in modelling and estimation, making it more favorable when either the parametric form of the model is unknown, or the model is incorrectly specified. An unconfounded assumption assessment test based on the estimated effects in the pre-treatment period and a normalized placebo test is developed to determine the significance of the estimated treatment effects. The proposed DSC method is demonstrated by numerical simulations and real data examples that highlight the effects of air pollution alerts in Beijing and the COVID-19 lockdown in Shanghai.

C1532: **Model-free data integration for estimation of average treatment effect in randomized clinical trial**
Presenter: **Kosuke Morikawa**, Iowa State University, United States

In randomized clinical trials, the enrollment of subjects is often limited by time and costs, hindering valid statistical analysis. One approach to address this issue is utilizing readily available published summary data. However, this method encounters challenges due to potential differences in background information between the current study and external data. A method is proposed that facilitates the integration of potentially biased summary data, including observational data, without necessitating additional models typically associated with regression or propensity score models.

CO002 Room K2.40 RECENT ADVANCES IN STATISTICAL METHODS FOR PRACTICAL APPLICATIONS Chair: Rui Pan

C0554: **Network tight community detection**
Presenter: **Huimin Cheng**, Boston University, United States

Conventional community detection methods often categorize all nodes into clusters. However, the presumed community structure of interest may only be valid for a subset of nodes (named tight nodes), while the rest of the network may consist of noninformative scattered nodes. For example, a protein-protein network often contains proteins that do not belong to specific biological functional modules but are involved in more general processes or act as bridges between different functional modules. Forcing each of these proteins into a single cluster introduces unwanted biases and obscures the underlying biological implication. To address this issue, a tight community detection (TCD) method is proposed to identify tight communities excluding scattered nodes. The algorithm enjoys a strong theoretical guarantee of tight node identification accuracy and is scalable for large networks. The superiority of the proposed method is demonstrated by various synthetic and real experiments.

C0616: **LMANStat: A multi-layer academic network dataset derived from statistical publications**
Presenter: **Rui Pan**, Central University of Finance and Economics, China

The utilization of multi-layer network structures now enables the explanation of complex systems in nature from multiple perspectives. Multi-layer academic networks capture diverse relationships among academic entities, facilitating the study of academic development and the prediction of future directions. However, there are currently few academic network datasets that simultaneously consider multi-layer academic networks; often, they only include a single layer. A large-scale multi-layer academic network dataset is provided, namely, LMANStat, which includes collaboration, co-institution, citation, co-citation, journal citation, author citation, author-paper and keyword co-occurrence networks. Furthermore, each layer of the multi-layer academic network is dynamic. Additionally, the attributes of nodes are expanded, such as authors' research interests, productivity, region and institution. Supported by this dataset, it is possible to study the development and evolution of statistical disciplines from multiple perspectives. This dataset also provides fertile ground for studying complex systems with multi-layer structures.

C0775: **Subsampling spectral clustering for stochastic block models in large-scale networks**
Presenter: **Danyang Huang**, Renmin University of China, China

The rapid development of science and technology has generated large amounts of network data, leading to significant computational challenges for network community detection. A novel subsampling spectral clustering algorithm is proposed to address this issue, which aims to identify community structures in large-scale networks with limited computing resources. The algorithm constructs a subnetwork by simple random subsampling from the entire network, and then extends the existing spectral clustering to the subnetwork to estimate the community labels for entire network nodes. As a result, for large-scale datasets, the method can be realized even using a personal computer. Moreover, the proposed method can be generalized in a parallel way. Theoretically, under the stochastic block model and its extension, the degree-corrected stochastic block model, the theoretical properties of the subsampling spectral clustering method are correspondingly established. Finally, to illustrate and evaluate the proposed method, a number of simulation studies and two real data analyses are conducted.

CO346 Room S0.03 MATHEMATICS OF DEEP-LEARNING FOR SOLVING DIFFERENTIAL EQUATIONS Chair: Anirbit Mukherjee

C0277: **Filtering dynamical systems using observations of statistics**
Presenter: **Eviatar Bach**, University of Reading, United Kingdom
Co-authors: Tim Colonius, Isabel Scherl, Andrew Stuart

The problem of filtering dynamical systems is considered, possibly stochastic, using observations of statistics. Thus, the computational task is to estimate a time-evolving density $\rho(v,t)$ given noisy observations of the true density $\rho^\dagger$; this contrasts with the standard filtering problem based on observations of the state v. The task is naturally formulated as an infinite-dimensional filtering problem in the space of densities ρ.

However, for the purposes of tractability, algorithms are sought in state space; specifically, a mean-field state-space model is introduced, and using interacting particle system approximations to this model, an ensemble method is proposed. The resulting methodology is referred to as the ensemble Fokker-Planck filter (EnFPF). Under certain restrictive assumptions, it is shown that the EnFPF approximates the Kalman-Bucy filter for the Fokker-Planck equation, which is the exact solution to the infinite-dimensional filtering problem. Furthermore, the numerical experiments show that the methodology is useful beyond this restrictive setting. Specifically, the experiments show that the EnFPF is able to correct ensemble statistics, to accelerate convergence to the invariant density for autonomous systems, and to accelerate convergence to time-dependent invariant densities for non-autonomous systems. Possible applications of the EnFPF are discussed in climate ensembles and in turbulence modeling.

C0881: **Measuring the risk of solving fluid dynamics by neural nets**
Presenter: **Dibyakanti Kumar**, The University of Manchester, United Kingdom
Co-authors: Anirbit Mukherjee

What is the relationship between a machine learning model's error in approximating the PDE solution and its physics-informed neural net (PINN) loss function value? This is critical for scientific-ML, and it is generally quite unclear. The attempt is to prove relationships between these two components for non-viscous pressure-less fluids (Burgers' PDE) in arbitrary dimensions while allowing for flow divergence. This is an interesting edge because it allows for PDE solutions that blow up in finite time while starting from smooth solutions - and hence, it allows for hard tests of theory to be conducted. Recently, some other empirical studies have pointed out that such solutions might be detectable by the PINN method if one regularizes for the gradients of the neural surrogate. Interestingly, this way of doing the population risk analysis reveals that this risk does indeed scale with such functional norms of the surrogate - and hence, it gives theoretical foundations for such penalizers to be added to PINN losses. Can risk bounds be obtained on PINNs that vanish in the limit of a large number of collocation points being used? That remains unclear - and definitely so for these cases as considered. Hence, the neural PDE solving scenario considered motivates exciting new research directions about generalization error bounding.

C1052: **Statistical learning theory for neural operators**
Presenter: **Jakob Zech**, Heidelberg University, Germany
Co-authors: Niklas Reinhardt, Sven Wang

Convergence rates for neural network-based operator surrogates are discussed, which approximate smooth maps between infinite-dimensional Hilbert spaces. Such surrogates have a wide range of applications and can be used in uncertainty quantification and parameter estimation problems in fields such as classical mechanics, fluid mechanics, electrodynamics, earth sciences, etc. The operator input represents the problem configuration and models initial conditions, material properties, forcing terms, and/or the domain of a partial differential equation (PDE) describing the underlying physics. The operator output is the corresponding PDE solution. The analysis demonstrates that, under suitable smoothness assumptions, the empirical risk minimizer for specific neural network architectures can overcome the curse of dimensionality in terms of required network parameters and the input-output pairs needed for training.

CO300 Room S0.11 STATISTICAL METHODS FOR HIGH-DIMENSIONAL NEUROIMAGING AND TIME SERIES DATA Chair: Ali Shojaie

C0724: **Instrumental variable analysis with multivariate point process treatments**
Presenter: **Shizhe Chen**, University of California, Davis, United States
Co-authors: Zhichao Jiang, Yu Liu

Multivariate point processes are popular tools for inferring relationships among subjects from recurrent event data such as neural spike trains. Complicated by the unmeasured confounding variables, interventions to the system are often employed in order to infer causality. However, these interventions are of low precision, and they might cause the intensity of multiple processes simultaneously. An instrumental variable framework is proposed, with treatments being multivariate point processes. It is shown that the causal effects can be learned using generalized Wald estimation. A penalized estimation procedure is proposed, motivated by classic methods for density deconvolution. The proposed method is applied to neural data from behavioral experiments on mice.

C0839: **Spectral differential network analysis for high-dimensional time series**
Presenter: **Michael Hellstern**, University of Washington, United States
Co-authors: Ali Shojaie, Byol Kim

Analyzing multivariate time series networks is popular in many fields, from neuroscience to seismology and signal processing. In particular, partial spectral coherence is a common choice of the network due to its representation as the frequency domain correlation between two variables after removing the best linear predictor of all other variables. In many applications, it is often of interest to study how these networks change across different conditions. For example, in neuroscience, one might be interested in how the brain network changes before and after stimulation. Estimates of differential networks typically rely on estimating the network in each condition and naively taking their difference. In high dimensions, establishing consistency of these estimates requires the restrictive assumption of the sparsity of each network. A direct estimator of the difference in inverse spectral densities of time series in two conditions is proposed. Using an L1 penalty on the difference, consistency is established only by requiring the difference to be sparse. This is a more realistic assumption if, for example, there are minimal differences in the networks between conditions. The difference estimator is further debiased to obtain asymptotically valid inference.

C1037: **Transfer learning for high-dimensional reduced rank time series models**
Presenter: **Abolfazl Safikhani**, George Mason University, United States

The objective of transfer learning is to enhance estimation and inference in target data by leveraging knowledge gained from additional sources. Recent studies have explored transfer learning for independent observations in complex, high-dimensional models assuming sparsity, yet research on time series models remains limited. The focus is on transfer learning for sequences of observations with temporal dependencies and a more intricate model parameter structure. Specifically, the vector autoregressive model (VAR), a widely recognized model for time series data, is investigated where the transition matrix can be deconstructed into a combination of a sparse matrix and a low-rank one. A new transfer learning algorithm is proposed and tailored to estimate high-dimensional VAR models characterized by low-rank and sparse structures. Additionally, a novel approach is presented for selecting informative observations from auxiliary datasets. Theoretical guarantees are established, encompassing model parameter consistency, informative set selection, and the asymptotic distribution of estimators under mild conditions. The latter facilitates the construction of entry-wise confidence intervals for model parameters. Finally, the empirical efficacy of the methodologies is demonstrated through both simulated and real-world datasets.

CO312 Room S0.12 ANALYSIS OF EXTREME VALUES: THEORY AND APPLICATIONS Chair: Boris Beranger

C1097: **Flexible max-stable processes for fast and efficient inference**
Presenter: **Peng Zhong**, University of New South Wales, Australia
Co-authors: Scott Sisson, Boris Beranger

Max-stable processes serve as the fundamental distributional family in extreme value theory. However, likelihood-based inference methods for max-stable processes still heavily rely on composite likelihoods, rendering them intractable in high dimensions due to their intractable densities. A fast and efficient inference method is introduced for max-stable processes based on their angular densities for a class of max-stable processes whose angular densities do not put mass on the boundary space of the simplex, which can be used to construct r-Pareto processes. The efficiency of the proposed method is demonstrated through two new max-stable processes, the truncated extremal-t process and the skewed Brown-Resnick process. The proposed method is shown to be computationally efficient and can be applied to large datasets. Furthermore, the skewed Brown-Resnick process contains the popular Brown-Resnick model as a special case and possesses nonstationary extremal dependence structures. The new max-stable processes are showcased on simulated and real data.

C1106: **Modeling high and low extremes with a novel dynamic spatiotemporal model**
Presenter: **Likun Zhang**, University of Missouri, United States
Co-authors: Christopher Wikle

In numerous dynamic systems, significant environmental challenges, including severe weather events and abrupt climate changes, have become prevalent. In order to fully understand the underlying mechanisms and enhance informed decision-making, a flexible model capable of accommodating extremes is necessary. The existing dynamic spatio-temporal models exhibit limitations in capturing extremes when assuming Gaussian error distributions, whereas the current models for spatial extremes are focused on joint upper tails at two or more locations while assuming temporal independence in the copula-based modeling framework. A novel class of dynamic spatiotemporal models are introduced, capable of accommodating both high and low extremes through a mixture of stable distributions with varying tail indices. Due to a redistribution kernel embedded in the hierarchical construct, the model can describe complex advective and diffusive dynamics with relatively few parameters and characterize differing levels of same-tail and opposite-tail extremal dependence, which are non-stationary across space and time. The effectiveness of the methods is demonstrated by applying them to turbulence flow observations that are chaotic and highly irregular.

C1148: **Bayesian inference for functional extreme events defined via partially unobserved processes**
Presenter: **Max Thannheimer**, University of Stuttgart, Germany
Co-authors: Marco Oesting

In order to describe the extremal behaviour of some stochastic process X, approaches from univariate extreme value theory are typically generalized to the spatial domain. A generalized peaks-over-threshold approach can be used, allowing the consideration of single extreme events. These can be flexibly defined as exceedances of a risk functional ℓ, such as a spatial average, applied to X. Inference for the resulting limit process, the so-called ℓ-Pareto process, requires the evaluation of $\ell(X)$ and thus the knowledge of the whole process X. In practical application, the challenge is that observations of X are only available at single sites. To overcome this issue, a two-step MCMC algorithm is proposed in a Bayesian framework. First, it is sampled from X conditionally on the observations to evaluate which observations lead to ℓ-exceedances. In the second step, these exceedances are used to sample from the posterior distribution of the parameters of the limiting ℓ-Pareto process. Alternating these steps results in a full Bayesian model for the extremes of X. It is shown that, under appropriate assumptions, the probability of classifying an observation as ℓ-exceedance in the first step converges to the desired probability. Furthermore, given the first step, the distribution of the Markov chain constructed in the second step converges to the posterior distribution of interest.

CO105 Room S0.13 UNCERTAINTY QUANTIFICATION Chair: David Woods

C0372: **Comparative analysis of model discrepancy treatment: Calibration versus scientific machine learning**
Presenter: **Victoria Volodina**, University of Exeter, United Kingdom

In healthcare and biological systems, mathematical models are increasingly used to understand complex biological processes. However, many of these models suffer from model inadequacy, posing a significant challenge to their use in clinical decision-making. Two main methods are reviewed to deal with model errors. Within the uncertainty quantification (UQ) community, model error is considered during calibration, where the observation is expressed as the sum of three terms: the simulator output at the true values of the calibration parameters, the model discrepancy, and the observation error. In calibration, a mathematical model is treated as a "black-box" system with the main objective of learning the values of calibration parameters. An alternative approach is to construct a hybrid "grey-box" model by filling in the incomplete parts of the computational model with a non-parametric model. The non-parametric model is used to learn the missing processes by comparing the available observations with the computational model output. To provide interpretability, the outputs of the non-parametric model are then regressed back down to symbolic form to learn the missing terms from the model using symbolic regression. These two methods are compared, and their performance is illustrated in an application to Siggaard-Andersen oxygen status algorithm.

C0625: **Linked deep Gaussian process emulation of model networks**
Presenter: **Deyu Ming**, University College London, United Kingdom

Computer models can be expensive and quickly become computationally prohibitive with large simulations. Statistical emulators, such as Gaussian processes, are thus essential for accelerating these simulations, enabling efficient downstream analyses like sensitivity analysis, calibration, and optimization, especially when computational resources are limited. Modern scientific problems often span multiple disciplines, necessitating the integration of distinct computer models. Building statistical emulators for such networks of computer models is challenging due to each model's unique functional complexities, computation times, and programming environments. Linked deep Gaussian process (LDGP) emulation offers a powerful solution by conceptualizing a computer model network as deep Gaussian processes with partially exposed hidden layers. Stochastic imputation, which integrates the expectation-maximization algorithm with elliptical slice sampling, is developed to infer these partially exposed deep networks. Synthetic and empirical examples, including the Joint UK Land environment simulator (JULES) and LDGP emulators, augmented by sequential designs and automatic structural pruning, have been found to perform significantly better than conventional Gaussian process emulators in predictive accuracy and uncertainty quantification. The implementations of these examples are facilitated by the dgpsi package, which is publicly available for both R and Python users on CRAN and CONDA.

C0928: **Bayes linear analysis with uncertain covariates**
Presenter: **Samuel Jackson**, Durham University, United Kingdom
Co-authors: David Woods

Statistical models typically capture uncertainties in existing knowledge of the corresponding real-world processes. However, it is less common for this uncertainty specification to capture uncertainty surrounding the values of the covariates to the model, which are often assumed to be known. General modelling methodology is developed with uncertain covariates in the context of the Bayes linear paradigm, which involves adjustment of second-order belief specifications over all the quantities of interest only, without the requirement for probabilistic specifications. In particular, an extension of commonly employed second-order modelling assumptions is proposed for the case of uncertain covariates, with explicit implementation in the context of regression analysis, stochastic process modelling, and statistical emulation. The methodology is demonstrated in the context of extracting aluminum by electrolysis and emulation of a complex network of functions.

CO014 Room BH (S) 1.01 Lec. Theatre 1 TIME SERIES ECONOMETRICS Chair: Antonio Montanes

C1458: **Drivers of portfolio equity and bond investment in the European Union: The interplay of tax havens and gravity factors**
Presenter: **Mariam Camarero**, University Jaume I, Spain

The purpose is to examine the determinants of portfolio equity and bond investment in the European Union. The impact of different drivers typical of the gravity model developed by another study is estimated. A notable aspect is that it accounts for the effects of tax havens through a recent database. Findings suggest that gravity variables (distance, economic size, and resistance), as well as historical links and global risk, explain portfolio holdings allocation. The model also captures the positive effect of government quality and financial development on portfolio equity and bonds. Given the differences in nature and risk between assets, the results are also compared for portfolio equity and bonds, finding that while portfolio equity is more mobile, portfolio debt tends to be invested in neighboring countries; more specifically, EU debt tends to remain in the EU. Results also suggest that portfolio equity is more affected by global risk and multilateral financial restrictions. Finally, the comparative analysis using the IMF CPIS database (constructed under the residence principle) shows that not accounting for tax havens underestimates the gravity and fundamental factors explaining portfolio equity and bonds holdings investment.

C1459: **Estimation of the autoregressive parameter under the presence of outliers**
Presenter: **Antonio Montanes**, University of Zaragoza, Spain

The purpose is to explore the properties of the estimator of the autoregressive parameter proposed in another study when the variable under analysis may exhibit some additive outliers. Results show that this statistic exhibits better size and power properties than the standard alternatives, especially when the autoregressive parameter is highly persistent.

C1614: **Residual-based tests for cointegration involving bounded stochastic processes**
Presenter: **Josep Lluis Carrion-i-Silvestre**, Universitat de Barcelona, Spain

The proposal is to cointegrate test statistics that account for the fact that some of the variables in the system are bounded. A generalized least squares-based estimation technique is implemented to improve the power properties of test statistics, which implies the use of a non-centrality parameter that considers the bounded nature of the variables.

CO202 Room BH (SE) 1.01 BAYESIAN ECONOMETRICS IN ECONOMICS AND FINANCE Chair: Aristeidis Raftapostolos

C1211: **Joint quantile shrinkage: Toward non-crossing Bayesian quantile models**
Presenter: **Tibor Szendrei**, Heriot-Watt University, United Kingdom
Co-authors: David Kohns

The prevalence of crossing quantiles has led to various methods for estimating monotonically increasing quantiles. Recent research has shown that one can achieve non-crossing quantiles by jointly estimating the quantiles and imposing fused shrinkage with quantile-specific hyperparameters. This approach is extended to the Bayesian realm. To achieve this, the Bayesian Quantile norm of using the Asymmetric Laplace distribution is deviated from, and instead, a general Bayes method is opted for to minimize a loss function. Additionally, instead of directly estimating quantiles, the focus is on estimating and shrinking the differences between quantiles. The formulation aligns with the time-varying parameter (TVP) models common in macroeconometrics, allowing us to use efficient TVP samplers for estimation. It is demonstrated that the proposed framework provides superior fits compared to conventional Bayesian and frequentist quantile regression estimators.

C1251: **Variable ordering in a Cholesky- multivariate stochastic volatility model**
Presenter: **Martina Zaharieva**, CUNEF Universidad, Spain
Co-authors: Ping Wu

Modeling the time-varying structure of the covariance matrix of financial time-series has been widely studied both theoretically and empirically in the finance literature. Recently, the Cholesky-decomposition-based approach to multivariate stochastic volatility has been established as a flexible and interpretable alternative. Despite that, the issue of order dependence in this type of model has not been the primary focus. The aim is to propose a prior for variable ordering in a Cholesky-type multivariate stochastic volatility model, where a natural ordering of the time series is not available. The approach to forecasting a large cross-section of stock returns is applied.

C1695: **Nuclear norm penalised non-linear modelling for asset pricing**
Presenter: **Aristeidis Raftapostolos**, Kings College London, United Kingdom
Co-authors: Ilias Chronopoulos, George Kapetanios

A nonlinear and non-parametric estimator for interactive fixed-effect panel regressions is proposed. Further, we extend the nuclear norm penalized estimation in a nonlinear setup using neural networks and apply our procedure to a large factor data library containing risk factors discovered in the literature in the last 30 years. Our estimator achieves robust out-of-sample performance. Most importantly, we move elements of model interpretability and explainability to the foreground.

CO047 Room BH (SE) 1.02 BAYESIAN COMPUTATION Chair: Michael Daniels

C1267: **Uncertainty quantification in latent position graph models**
Presenter: **Nick Heard**, Imperial College London, United Kingdom

From a graph-based perspective, anomaly detection techniques currently deployed in enterprise cyber-security typically act on individual nodes or edges, for example, tracking connectivity patterns of a network host over time or detecting unusual volumes or periodicity in data transfers between two network nodes. Techniques which leverage the full network graph are less common; global network models have typically proved too simplistic in their assumptions, such as the well-studied but arguably overused stochastic block model. A new anomaly detection framework is proposed, which seeks to fully quantify uncertainty in node positions for latent position network graph models. Such a framework admits the possibility for nodes to be identified as outlying through, for example, unusual entropy levels in their perceived graph position rather than simply relying on detecting spatial outliers.

C1366: **Efficient Bayesian inference on sparse and low-rank covariance matrices via projection**
Presenter: **Maoran Xu**, Indiana University, United States

In covariance estimation for high-dimensional data, sparsity and low-rank assumptions are commonly used to reduce dimensionality. However, in a Bayesian paradigm, it is challenging to conduct posterior computation under priors that simultaneously impose sparsity and low-rank (SLR) structure. To bypass the usual challenges inherent in computation for orthogonal and sparse matrix factorizations, a novel transformation-based approach is proposed. A normal-inverse-Gamma prior is projected from to the SLR space by thresholding the row-norm and trace norm, leading to a projected SLR (PSLR) prior. Remarkably, it is shown that it is possible to conduct posterior computation under the PSLR prior by sampling from the conjugate normal-inverse-Gamma posterior and projecting the draws. This dramatically simplifies computation. The resulting posterior distribution is shown to satisfy frequentist optimality properties, including adaptive posterior contraction rates. The approach is evaluated in simulation studies and applied to a gene expression data set.

C1489: **Variational Bayes and truncation approximations for enriched Dirichlet process mixtures**
Presenter: **Somnath Bhadra**, University of Florida, United States
Co-authors: Michael Joseph Daniels

A common impediment in conducting inference for Bayesian nonparametric models is either the need for complex MCMC algorithms and/or computational run-time for large datasets. Solutions are proposed for enriched Dirichlet process mixtures (EDPM). A variational Bayes estimator is derived based on a previously developed truncation approximation for EDPMs. The variational Bayes estimator can be used in two ways: 1. to develop a more efficient truncation approximation; 2. as good initial values for a blocked Gibbs sampler based on this more efficient truncation approximation or for a polya urn sampler. The accuracy of this more efficient truncation approximation is derived, and it is demonstrated how this allows for simple implementation of EDPMs in standard software for Hamiltonian MC. The validity of the approximations is confirmed by simulations.

CO185 Room BH (SE) 1.05 ECONOMETRICS AND FORECASTING FOR ECONOMIC AND FINANCIAL MARKETS Chair: Rustam Ibragimov

C1093: **Abrupt variance shifts and volatility forecasting in the renewable energy markets: A comprehensive analysis**
Presenter: **Akram Hasanov**, Monash University, Malaysia

Research on volatility modelling in the renewable energy markets has significantly increased over the past decade owing to the crucial role played by renewable energy sources in addressing climate change concerns. Using daily data on various renewable energy stock indices, various econometric models are comprehensively examined with and without accommodating structural shifts in the variance, and their forecasting performance is assessed at multiple horizons through a sliding window scheme. One of the novel aspects is the consideration of the forecasting capabilities of GARCH-class models, augmenting the regime dummies in the variance models. First, single regime GARCH-class and stochastic volatility models are used with different window sizes and distributional assumptions. Furthermore, MS-GARCH models are considered under several distribution assumptions. Second, a subset of competing models employs the break detection algorithm to determine the estimation windows considering the existence of structural breaks. The main results show that incorporating the endogenously detected structural breaks in the model estimations leads to considerable forecasting accuracy gains. The findings enrich the ongoing debate on the most effective approach to accounting for structural shift information to enhance the precision of volatility forecasts in the out-of-sample analysis of renewable energy stocks.

C1662: **Robust inference in predictive regressions for stock returns**
Presenter: **Rustam Ibragimov**, Imperial College Business School, CEBA and New Economic School, United Kingdom
Co-authors: Jenny Hau-Ruess, Aleksey Min

The focus is on a detailed robust analysis of predictability of stock returns. While a prior study suggests that no regression model can forecast the equity premium more accurately than its historical average, another study concludes that there are, in fact, such regressions if one restricts the model appropriately. The contribution to this discussion is by applying the general heterogeneity and autocorrelation robust inference approaches developed in other studies. These approaches are based on t-statistics computed using group estimates of regression parameters dealt with. They have appealing finite sample properties in many heterogeneity and dependence settings observed in practice as compared to, e.g., widely used HAC inference methods. The approaches are simple to use and do not require the estimation of consistent standard errors of regression estimators. The robust t-statistic methods are applied to assess the significance of the regressions slope coefficients, and the conclusions are compared with the results obtained from widely used alternative methods. It is examined whether a regression forecast is significantly more accurate as compared to the simple historical average forecast. Results demonstrate that even a critical investor can find a regression model that predicts the equity premium more accurately than the historical average based on data up to the year 2005.

C1056: **Stylized facts of cryptocurrency markets: Robust definitions and inference approaches**
Presenter: **Nursultan Abdullaev**, Innopolis University and Centre for Econometrics and Business Analytics SPbSU, Russia
Co-authors: Rustam Ibragimov

Several works in the literature have focused on the analysis of key stylised facts of cryptocurrency returns linked to fundamental problems of efficiency and predictability of cryptocurrency markets, including (i) heavy tails, indicating, in particular, that large price/return downfalls and fluctuations are more common than might be expected under a normal distribution; (ii) absence of autocorrelations, implying that return time series are to some extent are unpredictable and do not exhibit linear dependence over time; and (iii) volatility clustering, where periods of high volatility tend to be followed by similar periods and likewise for low volatility, implying nonlinear dependence in return time series. The contribution is the detailed study of the above properties of cryptocurrency markets using recently developed robust, econometrically and statistically justified definitions of and methods for inference on market (non-)efficiency, volatility clustering, and nonlinear dependence in return time series. In contrast to existing methodologies, the inference methods used in the analysis are robust against nonlinear dynamics and tail-heaviness of returns. The results of the analysis have significant implications for econometric modelling, risk management, and policy formulation in the context of cryptocurrency trading and investment.

CO106 Room BH (SE) 1.06 MODELLING INVESTMENT DECISION MAKING AND ASSET PRICING Chair: Michail Karoglou

C1057: **Hedge fund performance, classification with machine learning, and managerial implications**
Presenter: **Dimitrios Stafylas**, University of York, United Kingdom
Co-authors: Emmanouil Platanakis, Charles Sutcliffe, Wenke Zhang

Prior academic research on hedge funds focuses predominately on fund strategies in relation to market timing, stock picking, and performance persistence, among others. However, the hedge fund industry lacks a universal classification scheme for strategies, leading to subjective fund classifications and inaccurate expectations of hedge fund performance. Machine learning techniques are used to address this issue. First, it examines whether the reported fund strategies are consistent with their performance. Second, it examines the potential impact of hedge fund classification on managerial decision-making. Results suggest that for most reported strategies, there is no alignment with fund performance. Classification matters in terms of abnormal returns and risk exposures, although the market factor remains consistently the most important exposure for most clusters and strategies. An important policy implication of our study is that the classification of hedge funds affects asset and portfolio allocation decisions, as well as the construction of the benchmarks against which performance is judged.

C1058: **When Black-Litterman meets decision-fusion for asset allocation**
Presenter: **Emmanouil Platanakis**, University of Bath - School of Management, UK, United Kingdom
Co-authors: Xinyu Huang, David Newton, Xiaoxia Ye

Efficient decision-making often requires synthesizing insights from diverse knowledge domains. A novel method to model and consolidate inputs from multiple experts within the Black-Litterman model (BLM) is introduced. Each expert's beliefs and subjective judgements on stock trends are formulated as belief distributions in the decision fusion framework. Both misspecification and estimation errors are accounted for in the equilibrium allocations and covariance matrix. For equilibrium allocations in BLM, the mean-variance asset allocation rule is optimally combined with either the minimum-variance or the 1/N rule. Wishart stochastic volatility for modeling covariances is employed, initializing with a sparse conditional covariance matrix via the graphical lasso. After transaction costs, the method consistently boosts the Sharpe ratio compared to the 1/N rule and surpasses eight other established portfolio strategies for managing estimation risk. Robustness experiments show that the approach withstands various parameter choices.

C1078: **Empirical decisions and replicating anomalies: The benefit of the aggregate average**
Presenter: **Peng Li**, University of Bath, United Kingdom
Co-authors: Jiaqi Guo, George Korniotis, Alok Kumar

The reproducibility of asset pricing anomalies is examined by assessing the effects of empirical choices in constructing samples and portfolios. To reduce the influence of methodological variation, a two-stage bootstrapping procedure is proposed that accounts for sample period and methodological variability, and averages anomaly return spreads across methods. The aggregate average approach mitigates p-hacking and false discoveries. Analyzing 173 anomalies using 96 methods, it is found that while p-hacking is possible, particularly with a t-statistic threshold of 2, around 70% of the 173 anomalies can be successfully replicated.

CO336 Room BH (S) 2.01 ADVANCES IN FINANCIAL ECONOMETRICS AND RISK ANALYTICS Chair: Mike So

C0590: **Bayesian analysis of long memory and roughness in financial volatility**
Presenter: **Toshiaki Watanabe**, Hitotsubashi University, Japan
Co-authors: Jouchi Nakajima

Realized volatility (RV) calculated using intraday returns has recently been used as an accurate estimator of financial volatility. Some researchers have documented that the log-RV may follow a long-memory process, which is represented by a fractional Brownian motion with the Hurst exponent greater than 0.5 or a fractionally integrated process with a positive difference parameter. Recent studies show that the log difference in RV may be rough, which is represented by a fractional Brownian motion with the Hurst exponent less than 0.5 or a fractionally integrated process with a negative difference parameter. A discrete-time model that is consistent with these two phenomena is presented, and a Bayesian method is developed for the analysis of this model using Markov chain Monte Carlo. Empirical analysis using the RV of the Nikkei 225 stock index reveals that the long-term memory model has the best in-sample fit, and the model that takes into account both long-term memory and roughness has the highest volatility prediction accuracy, surpassing the heterogeneous autoregressive (HAR) model, which is known to have high volatility prediction accuracy.

C0666: **On "sandwich" variance estimation: Bayesian versus frequentist**
Presenter: **Cy Sin**, National Tsing Hua University, Taiwan

It is well known the Eicker-Huber-White variances are not only heteroskedasticity-robust and nonlinearity-robust but also nonnormality-robust. In a recent paper, some of the Eicker-Huber-White variances are reviewed. Among other things, they conclude: (a) Simulation studies suggest HC(4), a variant of robust variance estimator proposed by another study, does not over-reject or mildly under-rejects even in cases of non-normal distributions; (b) The original robust variance (denoted by HC(0) and its variants considered by the prevalent statistical software (such as R and STATA), are all asymptotically equivalent. The purpose is to take a Bayesian approach and consider the balanced loss function (BLF) proposed by a recent study. Unlike the conventional inference loss function (ILF), this function strikes a balance between estimation error and lack of fit. This function is, in turn, generalized upon what Zellner first proposed, which confines attention to normality-type likelihoods. The Bayesian estimator of the variance-covariance matrix is asymptotically equivalent to the frequentist estimator. Non-normal likelihoods are covered. Simulation studies that compare the Bayesian estimator with the conventional estimators are performed.

C0382: **Semi-parametric financial risk forecasting incorporating multiple realized measures**
Presenter: **Richard Gerlach**, University of Sydney, Australia
Co-authors: Rangika Peiris, Chao Wang, Minh-Ngoc Tran

A semi-parametric joint value-at-risk (VaR) and expected shortfall (ES) forecasting framework incorporating multiple realized measures is developed. The proposed framework extends the realized exponential GARCH model to be: i. Semi-parametrically estimated via a joint loss function; and ii. Allow a time-varying relationship between VaR and ES whilst further extending existing semi-parametric quantile time series models to incorporate multiple realized measures. A quasi-likelihood is built, employing the asymmetric Laplace distribution that is directly linked to a joint loss function, enabling Bayesian inference for the proposed model. An adaptive Markov chain Monte Carlo method is used for the model estimation. The empirical section evaluates the performance of the proposed framework with six stock markets from January 2000 to June 2022, covering the period of COVID-19. Three realized measures, including 5-minute realized variance, bi-power variation, and realized kernel, are incorporated and evaluated in the proposed framework. One-step-ahead VaR and ES forecasting results of the proposed model are compared to a range of parametric and semi-parametric models, lending support to the effectiveness of the proposed framework.

CO322 Room BH (S) 2.02 STATISTICS FOR COMPLEX-VALUED TIME SERIES AND INTEGER-VALUED TIME SERIES Chair: Yan Liu

C0552: **Expression of circular time series models with complex-valued stochastic processes**
Presenter: **Hiroaki Ogata**, Tokyo Metropolitan University, Japan

Circular data can be expressed by points on a unit circle of the complex plane. The proposal is to express circular time series models by complex-valued stochastic processes. Covariance and complementary covariance functions are first introduced. The Fourier transforms of them are called power and complementary power spectral density functions. By leveraging them, a periodic component of the circular time series data is extracted in the sense of the time axis. As examples of circular time series models, the circular mixture transition distribution model and the wrapped autoregressive model are seen. The former is explained to have the same autocovariance structure as that of the complex-valued autoregressive process, while the latter has non-vanishing autocovariance even when the lag h tends to be infinity.

C0927: **Recent developments in complex-valued and circular time series modeling**
Presenter: **Takayuki Shiohama**, Nanzan University, Japan
Co-authors: Hiroaki Ogata

Since an angular valued time series is expressed as a time series on a unit circle on the complex plane, parameter estimation techniques share some well-known methods in complex-valued data analysis. Recently, several fundamental properties of higher-order circular time series models have been developed, while there remain several problems in the parameter estimation for such data. Some topics related to circular and/or complex-valued time series modeling are introduced.

C1264: **Predictive inference for discrete-valued time series**
Presenter: **Maxime Faymonville**, TU Dortmund University, Germany
Co-authors: Carsten Jentsch, Efstathios Paparoditis

For discrete-valued time series, predictive inference cannot be implemented by constructing prediction intervals to some pre-determined coverage level. Although prediction sets rather than intervals respect the discrete nature of the time, they are generally unable to retain the desired coverage. To address this, reversing the construction principle is proposed by considering pre-defined sets of interest, and the conditional probability is estimated that a future observation falls in these sets, given the time series at hand. The accuracy of the corresponding prediction procedure is evaluated by quantifying the uncertainty associated with the estimation of the corresponding conditional probabilities. For this purpose, (non-)parametric approaches are considered, and asymptotic theory is derived for the estimators of the conditional probabilities focusing on the case of INAR and INARCH models. Since the established limiting distributions are typically cumbersome to apply in practice, suitable bootstrap approaches are proposed to evaluate the distribution of the estimators. Additionally, the problem of model misspecification is addressed, and a bootstrap is proposed based on the robustification of the procedure. Simulations investigate the finite sample performance of the methods developed by considering different (non-)parametric bootstrap implementations to account for the various sources of randomness and variability. Applications to real-life data sets also are presented.

CO158 Room BH (S) 2.03 FINANCIAL ECONOMETRICS: BUBBLES AND FORECASTING Chair: Robinson Kruse-Becher

C0222: **Bubbles and crashes: A tale of quantiles**
Presenter: **Efthymios Pavlidis**, Lancaster University Management School, United Kingdom

Periodically collapsing bubbles, if they exist, induce asymmetric dynamics in asset prices. The purpose is to show that unit root quantile autoregressive models can approximate such dynamics by allowing the largest autoregressive root to take values below unity at low quantiles, which correspond to price crashes, and above unity at upper quantiles, which correspond to bubble expansions. On this basis, two-unit root tests are employed based on quantile regressions to detect bubbles. Monte Carlo simulations suggest that the two tests have good size and power properties and can outperform recursive least-squares-based tests that allow for time variation persistence. The merits of the two tests are further illustrated in three empirical applications that examine Bitcoin, U.S. equity and U.S. housing markets. In the empirical applications, special attention is given to the issue of controlling for economic fundamentals. The estimation results indicate the presence of asymmetric dynamics that closely match those of the simulated bubble processes.

C0431: **Fundamentals of financial forecasting: Simplicity vs. complexity**
Presenter: **Dimitrios Thomakos**, National and Kapodistrian University of Athens, Greece

Forecasting of financial time series is a topic of broad interest to academics and industry practitioners alike. One of the purposes of financial forecasting is to construct quantitative trading strategies which are essentially attempts at market timing: The forward-looking decision of whether to buy or sell a financial asset. This original form of speculation, as old as the financial markets themselves, has a number of characteristics which make it amenable to either a simplistic or a complex form of analysis, for traders can use, for example, simple moving averages or complicated machine learning methods. A number of arguments are offered in favor of simplicity vs. complexity, and a number of suitable yet simple models are presented that are interpretable, understandable, easily computable and (backtested to be) profitable. If one can devise such forecasting models based on simple concepts that are fundamental to financial forecasting, then the need or usefulness of large and complicated models for this kind of academic or professional activity is called for a re-examination.

C0309: **The everything bubble: What is behind the surge in US equity prices**
Presenter: **Christoph Wegener**, Leuphana University Lueneburg, Germany
Co-authors: Tobias Basse, Michael Lamla, Stefano Maiani

The purpose is to investigate the recent surge in United States (US) equity prices, documenting the explosive growth of the Standard & Poor's 500 Index since 2009. Similar trends across various US equity market sectors and other US dollar-denominated assets suggest a potential widespread speculative bubble. Testing for co-explosiveness reveals a common trend between the volume of the Federal Reserve's balance sheet and US asset prices, linking monetary policy to the asset price surge. US policymakers have lowered long-term interest rates, including the risk-free rate, to artificially low levels. It is theoretically argued that a rational bubble is unlikely in such an environment. Instead, a bond market bubble, driven by the Federal Reserve's substantial purchases of fixed-income securities following the drop in short-term interest rates to nearly zero, offers a more compelling explanation for the equity price surge. This argument is supported by an empirical testing procedure against rational bubbles, which is based on expectations rather than the dividend stream. This approach accounts for anticipated interest rate levels, providing an advantage. Since the procedure does not indicate a rational bubble and explosive behavior is not inferred in dividend streams, the (artificially) low interest rates remain the most plausible explanation for the explosive equity price behavior.

CO275 Room BH (S) 2.05 J-ISBA SESSION ON RECENT ADVANCES IN BAYESIAN STATISTICS Chair: Andrea Cremaschi

C0822: **Bayesian methods for regression with confounding variables**
Presenter: **Luke Travis**, Imperial College London, United Kingdom

Bayesian methods are developed for regression models in the presence of unobserved confounding variables. Confounded versions of both the sparse high-dimensional linear regression and nonparametric regression settings are considered. Under suitable conditions on the prior, the Bayesian posterior contracts to the truth at the same rate, with no confounding, and which has the same order as the L1-loss provided by frequentist methods. Moreover, it is illustrated that for a specific choice of prior covariance, the spectral obtained transforms the proposed by a frequentist counterpart. The strong performance of the Bayesian method (and a computationally scalable variational approximation) is demonstrated in terms of estimation and model selection in a variety of different scenarios. Moreover, it is shown that the out-of-the-box uncertainty quantification provided by the posterior is reliable.

C0883: **Bayesian nonparametric mixture models and clustering for ecological risks**
Presenter: **Louise Alamichel**, Universite Grenoble Alpes, France
Co-authors: Julyan Arbel, Guillaume Kon Kam King, Igor Pruenster

Bayesian nonparametric mixture models are common for modeling complex data. These models are well-known for being consistent when used for density estimation. However, the consistency of the posterior distribution does not provide asymptotic guarantees in the context of clustering problems. Until recently, there has been a lack of asymptotic guarantees regarding the posterior number of clusters for these models. After studying the asymptotic properties of these models, one of them is applied to a real-world problem in ecotoxicology. A Bayesian nonparametric mixture model is proposed to assess the ecological risks of water contaminants. The choice of a Bayesian nonparametric approach offers several advantages, including its efficiency in handling small datasets typical of environmental risk assessments, its ability to provide uncertainty quantification, and its capacity for simultaneous density and clustering estimation. Through systematic simulation studies and analysis of real datasets, the superiority of the Bayesian nonparametric approach is demonstrated over classical methods for this problem.

C0924: **Vecchia Gaussian processes: Probabilistic properties and Bayesian nonparametrics**
Presenter: **Yichen Zhu**, Bocconi University, Italy
Co-authors: Botond Szabo

Gaussian processes are widely used to model spatial dependency in geostatistical data, yet the exact computation suffers an intractable time complexity of $O(n^3)$. Vecchia approximation has become a popular solution to this computational issue, where spatial dependency is characterized by a sparse directed acyclic graph (DAG) that allows scalable Bayesian inference. Despite the popularity in practice, little is understood about the Vecchia Gaussian processes themselves, let alone their theoretical guarantees when employed in a regression model. The probabilistic properties of Vecchia Gaussian processes are systematically studied when the mother Gaussian process is a Matern process. Under minimal regularity conditions and appropriate selection of the DAG, the Vecchia Gaussian process retains many desirable properties of the mother Gaussian process. These probabilistic properties further allow the development of Bayesian nonparametric theory for the Vecchia Gaussian process, where minimax optimality is obtained when either prior smoothness matches the posterior or in an adaptive hierarchical Bayesian setting.

CO133 Room BH (SE) 2.01 Y-SIS - ADVANCES IN STATISTICAL LEARNING AND CLUSTERING Chair: Carlo Zaccardi

C0404: **Understanding corporate default: The role of accounting and market information with a cluster-based matching procedure**
Presenter: **Alessandro Bitetto**, University of Pavia, Italy
Co-authors: Michele Modina, Stefano Filomeni

Recent evidence highlights the importance of hybrid credit scoring models in evaluating borrowers' creditworthiness. However, the current hybrid models neglect to consider the role of public-peer market information in addition to accounting information on default prediction. Novel evidence is provided on the impact of market information in predicting corporate defaults for unlisted firms. A sample of 10,136 Italian micro-, small-, and mid-sized enterprises (MSMEs) is employed that borrow from 113 cooperative banks from 2012-2014 to examine whether market pricing of public firms adds additional information to accounting measures in predicting the default of private firms. Specifically, the probability of default (PD) of MSMEs is estimated using equity price of size-and industry-matched public firms, and then advanced statistical techniques based on the parametric algorithm (multivariate adaptive regression spline) and non-parametric machine learning model (random forests) are adopted. Moreover, by using

Shapley values, the relevance of market information in predicting corporate credit risk is assessed. Firstly, the predictive power of Merton's PD on default prediction is shown for unlisted firms. Secondly, the increased predictive power of credit risk models that consider both the Merton's PD and accounting information to assess corporate credit risk is shown.

C0876: **Spherical double k-means**
Presenter: **Emiliano Seri**, University of Rome Tor Vergata, Italy
Co-authors: Ilaria Bombelli, Maurizio Vichi, Stella Iezzi

The spherical double k-means (SDKM) clustering method is introduced for text data. A novel approach for simultaneous clustering of terms and documents. Using the strengths of k-means, double k-means, and spherical k-means, SDKM addresses the challenges of high dimensionality, noise, and sparsity inherent in text analysis. The choice of the number of clusters is addressed, both for the words and documents, using the cluster validity index pseudo-F, and the reliability of the method is verified through simulation studies. SDKM is applied to the corpus of US presidential inaugural addresses, spanning from George Washington in 1789 to Joe Biden in 2021. The analysis reveals distinct clusters of words and documents that correspond to significant historical themes and periods, showcasing the method's ability to facilitate a deeper understanding of the data. Findings demonstrate the efficacy of SDKM in uncovering underlying patterns in textual data.

C0982: **Contrastive learning: A statistical approach**
Presenter: **Luca Scaffidi Domianello**, University of Catania, Italy
Co-authors: Salvatore Ingrassia

Contrastive or self-supervised learning is an intuitive learning principle alternative to the likelihood-based one, usually employed to estimate unnormalized models, that is models for which the density is not normalized. It has wide applicability in different statistical areas, one of the most important being density estimation. The latter, which is related to unsupervised learning, can be reformulated through logistic regression modeling as a supervised learning task. This approach has been largely developed in the area of machine learning in different fields like natural language processing and image modeling, just to name a few. Nevertheless, from a statistical point of view, this approach has attracted very little attention so far. The main idea of contrastive learning is to learn to classify between the data of interest and some artificially generated ad hoc reference, also called noise data, by contrasting them. In the present contribution, a detailed statistical framework of contrastive learning approach and large numerical studies are provided, illustrating the properties of the learning principle. As expected, the reference distribution affects the parameter estimates related to the interest data, then the choice of the reference distribution is a crucial task that needs to be further analyzed. A final application on real data is presented for conclusion.

CO284 Room BH (SE) 2.05 MODELLING AND ANALYSING THE TRANSITION TO GREEN ENERGY Chair: Stefan Wrzaczek

C0620: **Transition out of coal: A model based framework**
Presenter: **Stefan Wrzaczek**, International Institute for Applied Systems Analysis (IIASA), Austria
Co-authors: Dieter Grass, Michael Kuhn, Alexia Prskawetz, Omkar Patange

Coal is a major contributor to anthropogenic carbon emissions and climate change. Coal mining and combustion are also leading causes of premature mortality due to local air pollution. On the other hand, coal is central to many economies that rely on mining, transportation, energy production and exports. With the changing climate and rapidly depleting carbon budgets, the urgency for coal phase-out has become more prominent, and many economies are under pressure to transition away from coal in a time-bound manner. Since the coal phase-out is a medium- or long-term process, answering the research questions requires a dynamic consideration. Optimal pathways for a coal phase-out are studied within a small open regional economy consisting of a coal extraction sector, an energy sector composed of both coal-based and renewable power generation, final goods and a household sector that relies on coal and electricity. The conditions under which coal extraction and fossil power generation are phased out are studied depending on preferences, cost and price structures. A contribution is envisaged to the modelling framework to analyze regional coal phase-out policies and provide a systematic analysis of the dynamic processes associated with the social, economic, political and technical coal transitions in coal-based regional economies.

C1066: **On green reputation**
Presenter: **Francisco Cabo**, Universidad de Valladolid, Spain
Co-authors: Mercedes Molpeceres-Abella

The dynamic interaction between a standard brown firm and a green firm with less pollution, though more costly production technology, is analyzed. Consumers are willing to pay a premium price for green goods, whose size depends on the green firm's reputation. Reputation is measured by a state variable that grows with the technology gap between green and brown production, as well as with the green firms' advertising, representing green-washing. Production activities generate emissions accumulated as a pollution stock. The brown producer only determines the produced quantity, while the green producer also has to determine the level of green technology and the investment in green advertising. The solutions are characterized and compared with and without the possibility of green washing, and the role played by the sensitivity of reputation to advertising efforts is analyzed.

C1165: **Marginal emission factors in electricity markets: The role of renewable energy sources and electricity trade**
Presenter: **Luigi Grossi**, University of Parma, Italy
Co-authors: Filippo Beltrami, Mario Liebensteiner

The purpose is to measure how cross-border commercial exchanges of electricity influence carbon emissions in both the generating and the receiving country. This central question encompasses the study of how electricity trading between countries affects the overall carbon emissions of the involved countries. The aim is to build on these insights by providing updated estimates of MEFs in Germany, taking into account physical power trade with neighboring countries. In particular, the impact that the export of electricity generated in Germany has on domestic CO2 emissions and on emissions in France, Germany's main trading partner, is estimated. It is demonstrated that the corresponding parameter can be considered an MEF coefficient. This analysis will be conducted through a two-stage econometric model. In the first stage, the generation mix (including nuclear and renewable energy sources) in neighboring countries is examined to determine how it affects Germany's net electricity exports. The results of this first stage are then included in the second stage to address the endogeneity of exports' effect on carbon emissions. This approach allows the consistent estimation of the impact of net electricity exports (adjusted for generation sources) on Germany's CO2 emissions, considering demand, supply, and prices of CO2 and natural gas.

CO308 Room BH (SE) 2.09 ADVANCES IN MACROECONOMETRICS Chair: Danilo Leiva-Leon

C0455: **Theory coherent shrinkage of time-varying parameters in VARs**
Presenter: **Andrea Renzetti**, Bocconi University, Italy

A novel theory of coherent shrinkage prior is introduced for time-varying parameters VARs. The proposed prior can be used to sharpen inference about the time-varying parameters by leveraging on prior information from an underlying economic theory about the macroeconomic variables in the model. Exploiting prior information from conventional economic theory is revealed to form a prior for the time-varying parameters and significantly improves inference precision and forecast accuracy over the standard TVP-VAR. More specifically, using the classical 3-equation New Keynesian block to form a prior for the TVP-VAR substantially enhances forecast accuracy of output growth and of the inflation rate in a standard model of monetary policy. Additionally, prior information from economic theory can be used to address the inferential challenges faced by the standard TVP-VAR during the zero lower bound.

C1034: **Dancing with the R-star: Information Shocks in the "New Normal"**
Presenter: **Giulia Martorana**, Catholic University of the Sacred Heart, Italy
Co-authors: Efrem Castelnuovo, Aristotelis Margaris

Benigno, Hofmann, Nuo, and Sandri (2024) hypothesize that monetary policy shocks may be behind changes in the real natural interest rate, i.e., R-star. We quantify the impact of information shocks - news about future economic conditions released by the Federal Reserve when announcing its policy decisions - on R-star via an internal instruments VAR approach. We find a statistically significant, economically relevant, and substantially persistent positive impact of expansionary information shocks on R-star. We document evidence consistent with an innovation channel at work transmitting such shocks to the natural interest rate. Working with a longer sample that accounts for pandemic data and a second zero lower-bound episode (on top of the great recession's), we document a significant sign asymmetry, i.e., the business cycle and R-star respond more (less) strongly to information shocks associated with monetary policy easing (tightening). This empirical finding survives the inclusion in the VAR of a variety of macroeconomic indicators that account for the various shocks that hit the US economy during and in the aftermath of the pandemic, including fiscal stress, geopolitical risk, energy disturbances, and global supply chain pressures.

C1045: **On the effect of monetary policy shocks on the neutral rate of interest**
Presenter: **Danilo Leiva-Leon**, European Central Bank, Germany
Co-authors: Rodrigo Sekkel, Luis Uzeda

An econometric framework is presented designed to simultaneously estimate the neutral rate of interest while assessing the impact of monetary policy shocks on it. The proposed model is a time-varying parameter Bayesian VAR with stochastic volatility where the structural shocks identified within the model are allowed to potentially influence the evolving parameters. Using an identification strategy based on instrumental variables, it is shown that a monetary policy shock has a small but persistent impact on the real neutral rate of interest, which we model as a component of the long-run Fisher equation relationship.

CO182 Room BH (SE) 2.10 ANALYSIS OF COMPLEX MULTIVARIATE DATA Chair: Thomas Verdebout

C1457: **Autocalibrated predictors in nonlife insurance pricing**
Presenter: **Julien Trufin**, Universita Libre de Bruxelles, Belgium
Co-authors: Michel Denuit

By exploiting massive amounts of data, machine learning techniques provide actuaries with predictors exhibiting high correlation with claim frequencies and severities. However, these predictors generally fail to achieve financial equilibrium and thus do not qualify as pure premiums. Autocalibration effectively addresses this issue since it ensures that every group of policyholders paying the same premium is, on average, self-financing. Balance correction has been proposed as a way to make any candidate premium autocalibrated with the added advantage that it improves out-of-sample Bregman divergence and, hence, predictive Tweedie deviance. The aim is to prove that balance correction is also beneficial in terms of concentration curves and conditions being derived, ensuring that the initial predictor and its balance-corrected version are ordered in Lorenz order. Finally, criteria are proposed to rank the balance-corrected versions of two competing predictors in the convex order.

C1485: **Inference for nearly directional data**
Presenter: **Thomas Verdebout**, Universite Libre de Bruxelles, Belgium

Models for noisy directional data are considered, in which a radial noise makes the observations deviate from their theoretical hyperspherical sample space, namely a hypersphere centered at and with radius. Inference hypothesis testing, point estimation, and confidence zone estimation are considered on the location parameter. Several asymptotic scenarios are introduced, in which the radius of the hypersphere and, most importantly, the noise magnitude may depend on the sample size in an essentially arbitrary way. This allows for considering very diverse cases in which the prior information that the data belongs to a hypersphere is more and more, or on the contrary, less and less relevant. The investigation is based on Le Cams asymptotic theory of statistical experiments and aims at a full understanding of the resulting limiting experiments.

C1492: **On runs tests for directional data and their local and asymptotic optimality properties**
Presenter: **Maxime Boucher**, Universite de Namur (UNamur), Belgium
Co-authors: Thomas Verdebout, Yuichi Goto, Christian Francq

The problem of detecting some randomness within directional data is studied. Motivated by a real data example involving sunspot locations, a concept of runs properly adapted to the directional context is defined. It is then shown that tests based on the later runs enjoy some local and asymptotic properties against local alternatives with serial dependence. The finite sample performances of the tests are computed using Monte Carlo simulations, and their usefulness is shown in a real data illustration that involves the analysis of sunspots on the Photosphere.

CC470 Room S-2.25 METHODOLOGICAL STATISTICS Chair: Maria Brigida Ferraro

C1562: **On testing random effects in linear and non-linear mixed effects models**
Presenter: **Germaine Uwimpuhwe**, Durham University, United Kingdom
Co-authors: Reza Drikvandi

Linear and non-linear mixed models (LMMs and NLMMs) are frequently used to analyse longitudinal data with nested structures. Unlike LMMs, testing the necessity of random effects in NLMMs remains challenging due to the limitations of traditional likelihood-based methods, which are constrained by assumptions of boundary issues and normality, as well as convergence issues and difficulties with correlated random effects. To address those issues, a permutation-based test is developed and extended, originally designed for linear models, to the non-linear context. The approach uses non-parametric estimation techniques, including variance least squares (VLS) and method of moments (MM) based on two-stage and first-order approximation. Through extensive empirical analyses, these tests are compared with the traditional mixture of chi-square tests (MLRT) based on empirical power, Type I error rate, and bias. Results show that permutation-based tests outperform the traditional methods. For example, a test based on VLS achieves power between 92% and 100%, while MM-based methods range from 88% to 100%, with a controlled Type I error rate. Tests have stable convergence rates across scenarios, in contrast to the MLRT method, which suffers from convergence issues. The R package TestREnlme is developed for the implementation of these methods, facilitating non-parametric estimation and permutation tests for random effects.

C1347: **Novel constraints in empirical likelihood for ranked set sampling**
Presenter: **Soohyun Ahn**, Ajou University, Korea, South

An innovative empirical likelihood (EL) method tailored for ranked set sampling (RSS) is introduced, which effectively harnesses the inherent ranking information within the sampling process. The approach imposes a novel constraint, ensuring that the sum of the within-stratum probabilities for each rank stratum equals $1/H$, where H represents the number of rank strata. This constraint streamlines the application of EL across both balanced and unbalanced RSS, eliminating the need for subjective weight selection in unbalanced cases. The efficacy of the method is demonstrated through its application to one-sample mean testing, supported by a comprehensive numerical study and analysis of two real-world datasets, body fat data and dental size data.

C1417: **Model fitting using partially ranked data**
Presenter: **Mayer Alvo**, University of Ottawa, Canada
Co-authors: Xiuwen Duan

The importance of models for complete ranking data is well-established in the literature. Partial rankings, on the other hand, naturally arise when the set of objects to be ranked is relatively large. Partial rankings give rise to classes of compatible order, preserving complete rankings. An exponential model is defined for complete rankings, and it is calibrated on the basis of a random sample of partial rankings data. It appeals to the EM algorithm. The approach is illustrated in some simulations and in real data.

CC497 Room S-1.06 BAYESIAN CAUSAL INFERENCE Chair: Michael Schweinberger

C0782: **Bayesian ensemble learning for principal causal effects**
Presenter: **Chanmin Kim**, SungKyunKwan University, Korea, South
Co-authors: Corwin Zigler

Principal stratification analysis assesses how the causal effects of a treatment on a primary outcome vary across different strata of units, which are defined by their treatment effect on an intermediate variable. This task is significantly complicated when the intermediate variable is continuous, resulting in an infinite number of basic principal strata. To address this, a Bayesian nonparametric approach is used to flexibly evaluate treatment effects across these strata. The approach employs Bayesian causal forests (BCF), which simultaneously specify two Bayesian additive regression tree models: one for principal stratum membership and one for the outcome conditional on principal strata. BCF's ability to capture treatment effect heterogeneity is particularly useful for assessing treatment effects across the continuum of continuously scaled principal strata. Additionally, BCF offers benefits in targeted selection and regularization-induced confounding. The effectiveness of the method is shown through a simulation study, and this methodology is applied to investigate how the causal effects of power plant emissions control technologies on ambient particulate pollution vary with the technologies' impact on sulfur dioxide emissions.

C1484: **Bayesian estimation of causal effects from multiple datasets using structural causal models**
Presenter: **Shunsuke Horii**, Waseda University, Japan
Co-authors: Yoichi Chikahara

A novel Bayesian approach is proposed for estimating causal effects from experimental data with multiple interventions. The method builds on structural causal models (SCMs) by calculating the posterior distributions of the parameters of structural equations derived from multiple datasets. Using these posterior distributions, the posterior distributions of the causal estimands are computed. Unlike traditional approaches that estimate only the average causal effect (ACE) from multiple data sources, this method offers two key advantages. First, it achieves Bayesian optimality, leading to improved estimation performance across a range of scenarios. Second, the approach extends beyond the estimation of ACE to allow for the estimation of modified causal effects, such as those conditioned on specific covariates or interventions. This added flexibility enhances the interpretability and applicability of causal estimates in real-world settings where effect modification is crucial. Through experiments on simulated data, it is demonstrated that the Bayesian framework outperforms existing methods in terms of estimation accuracy and provides more nuanced insights into causal relationships.

C0235: **A Bayesian model for estimating spillover effects in unconventional monetary policy**
Presenter: **Andrea Mercatanti**, Sapienza University of Rome, Italy
Co-authors: Sharmistha Guha, Taneli Makinen

The focus is on a causal inference scenario where interference between units arises from equilibrium effects in markets. These are situations in which policy interventions, through modifications in supply and/or demand functions, influence equilibrium prices and consequently have an impact also on non-eligible units. The specific focus lies on equilibrium effects in bond markets resulting from an unconventional monetary policy implemented by the European Central Bank (ECB) between 2016 and 2018. This policy is known as the Corporate Sector Purchase Program, which aimed to further strengthen the pass-through of the Eurosystems asset purchases1 to the financing conditions of the real economy. The core idea of the proposal is to utilize data from markets that are likely unaffected by the policy under consideration, specifically the markets of Hong Kong and South Korea. This data can serve as additional controls, namely as observations that are not influenced by the policy either directly or indirectly. For inference, a Bayesian causal factor model is relied on where Bayesian shrinkage techniques are taken advantage of to simultaneously perform model selection and parameter estimation. In particular, the Bayesian Lasso shrinkage mechanism is employed after reparametrizing subject-specific factor loadings. Findings suggest that the Program had a statistically and economically significant negative impact on yield spreads for both the eligible and non-eligible bonds.

CC502 Room K0.16 COMPUTATIONAL AND METHODOLOGICAL STATISTICS Chair: Jan Gertheiss

C1718: **Improving finite sample estimates of principal components for high-dimensional data**
Presenter: **Nuwan Weeraratne**, University of Waikato, New Zealand
Co-authors: Lynette Hunt, Jason Kurz

Principal Component Analysis (PCA) is a method of compressing many data into a format that captures the essence of the original data. Moreover, PCA is a matrix decomposition technique that uses eigendecomposition. It quantifies variable relationships using covariance matrices, determines the data distribution, and assesses direction significance using eigenvalues. Therefore, the variance-covariance estimation is crucial for PCA. However, the traditional Maximum Likelihood Estimation (MLE) of covariance is asymptotically unbiased, but it gives poor estimates of the principal components and poorly conditioned estimates of the covariance matrix in high dimensional settings where the number of variables (p) exceeds the number of observations (n). To address the issues with PCA, we proposed a novel covariance estimation called the Pairwise Differences Covariance (PDC) estimation with four regularized versions of PDC (i.e., Standardized PDC (SPDC), Local Scaled PDC (LPDC), Scaled by Maximum Absolute Value PDC (MAXPDC), and Scaled by Range PDC method (RPDC)). In empirical comparisons with MLE and its existing most famous alternative, Ledoit-Wolf, the SPDC and all other regularized PDC estimators perform well in estimating the variance-covariance structure and principal components while minimizing the PCs overdispersion and cosine similarity error (CSE). Real data applications are presented.

C1719: **Computationally efficient sparse sufficient dimension reduction via least squares svm and its extensions**
Presenter: **Jungmin Shin**, Korea University, Korea, South
Co-authors: Seung Jun Shin

Sparse sufficient dimension reduction (SDR) is explored through the framework of the penalized principal machine (PM). We proposed the penalized principal least square support vector machine (P2LSM) as a primary example of this approach. The P2LSM employs a squared loss function, facilitating efficient computation using the group coordinate descent algorithm. Our method enhances computational efficiency compared to conventional sparse SDR methods, particularly for large-scale datasets. We also extend our approach to include penalized principal asymmetric least squares regression (P2AR), penalized principal L2-SVM (P2L2M), and penalized principal (weighted) logistic regression (P2(W)LR), demonstrating its versatility. Additionally, the computational advantage and oracle property of the proposed methods are investigated. Extensive simulations and real data analyses illustrate the efficacy of the proposed methods in yielding sparse, interpretable solutions without compromising predictive accuracy.

C1704: **Dependence maps: A graphical tool for visualizing multivariate dependence**
Presenter: **Ivan Medovikov**, Brock University, Canada

A graphical tool for visualizing complex relationships between the variables is demonstrated that is based on the difference of empirical copulas of the underlying variables that we call a "dependence map". We also present a version of the tool designed to assess comonotonicity of vectors, that is - to highlight vector dependence. We demonstrate the applicability of the tool by visually studying the dependence between hourly wholesale electricity prices and grid features such as load and weather in New York. We demonstrate how the tool can uncover instances of non-linear (e.g. tail) dependence.

CC413 **Room K2.41** **SURVIVAL ANALYSIS** **Chair: Kathrin Moellenhoff**

C1625: **Testing the effect of multiple covariates on the cure rate using martingale difference correlation**
Presenter: **Blanca Monroy-Castillo**, Universidade da Coruna, Spain
Co-authors: M Amalia Jacome, Ricardo Cao

In survival analysis, there are situations in which not all subjects are susceptible to the final event. Methods are now well-developed to deal with this kind of data, generally known as cure model analysis. Mixture cure models allow for the estimation of the probability of cure and the survival function for uncured subjects. One of the main goals when working with these types of models is to determine if the covariates have any effect on the cure rate. In the nonparametric sense, a test was proposed in a past study. To propose a new test, the focus is on a new measure. Distance correlation, proposed by another study, has been extensively studied and extended. One of these extensions is the martingale difference correlation, which measures the departure from conditional mean independence between a scalar response variable V and a vector predictor variable U. Moreover, martingale difference correlation and its empirical counterpart inherit several desirable features from distance correlation and sample distance correlation. A new significance test for multiple covariates is proposed using martingale difference correlation and its extensions.

C1429: **Modeling the restricted mean survival time as a function of time horizons with pseudo-value regression trees**
Presenter: **Alina Schenk**, University of Bonn, Germany
Co-authors: Matthias Schmid

The restricted mean survival time (RMST) has become an increasingly important estimand in time-to-event studies. Defined as the restricted area under the survival function over $[0,\tau]$, the RMST represents the average event-free survival time up to the time horizon τ. In practice, directly modeling the RMST conditional on covariates is particularly useful for assessing the impact of a treatment or exposure on the expected lifetime. However, most direct modeling approaches for the RMST focus on a single fixed time horizon $\tau > 0$. Choosing an appropriate value for τ can be challenging and has been widely discussed in the literature. The purpose is to introduce an alternative approach to modeling the RMST as a function of τ using pseudo-value regression trees (PRT). PRT are characterized by a multivariate regression tree built on a pseudo-value outcome and by successively fitting a set of regularized additive models to the data in the nodes of the tree using gradient boosting. Like previously published approaches, PRT models RMST values at various time horizons simultaneously and incorporates time-varying covariate effects. A simulation study and a real-world application are presented to demonstrate the properties of the proposed method.

C0361: **Estimation in a three-state model with interval-censored data**
Presenter: **Luis Machado**, University of Minho, Portugal
Co-authors: Marta Azevedo, Gustavo Soutinho

In many fields, including medical research, engineering, and the social sciences, analyzing time-to-event data is essential for uncovering underlying processes and facilitating decision-making. A common challenge in this analysis arises when events are confirmed to have taken place within specific time intervals, yet the exact timing within those intervals remains unknown, a phenomenon known as interval censoring. The focus is on a three-state progressive multi-state survival model where the intermediate state and/or the final state may be interval-censored. The primary aim is to estimate state occupation probabilities, which are crucial for understanding the dynamics of state transitions over time. Additionally, the estimation of the bivariate distribution of the gap times is considered. New estimation methods are introduced based on the Turnbull estimator of survival to address the challenges posed by interval-censored events. Imputation-based methods are also explored for estimating event times within the interval, such as using the midpoint, left-point, and right-point of the interval. Findings contribute to filling the gap in the literature regarding interval-censored multi-state models, providing valuable insights for researchers and practitioners dealing with such data.

CC431 **Room Safra Lec. Theatre** **FUNCTIONAL DATA ANALYSIS** **Chair: Alessia Pini**

C1300: **Additive regression for Riemannian functional responses**
Presenter: **Germain Van Bever**, Universite de Namur, Belgium
Co-authors: Jeong Min Jeon

Additive regression is explored for a functional response whose values lie on a general Riemannian manifold. Euclidean predictors that may not be directly observable but are estimable are also addressed, such as component scores obtained from dimension reduction. The smooth backfitting method is employed to estimate the additive model, and its asymptotic properties are derived. Additionally, novel dimension reduction techniques are discussed for general predictors in the presence of the Riemannian functional response. The usefulness of the approach is demonstrated through a real data application.

C1523: **A fully functional approach for statistical shape analysis**
Presenter: **Issam-Ali Moindjie**, University of Quebec in Montreal, Canada
Co-authors: Marie-Helene Descary, Cedric Beaulac

The shape $\tilde{\mathbf{X}}$ of a random planar curve, $\mathbf{X}$, is what remains when the deformation variables (scaling, rotation, translation, and parametrization) are removed. Previous studies in statistical shape analysis have focused on analyzing $\tilde{\mathbf{X}}$ through discrete observations of $\mathbf{X}$. While this approach has some computational advantages, it overlooks the continuous nature of variables: $\tilde{\mathbf{X}}$, $\mathbf{X}$, and it ignores the potential dependence of deformation variables on each other and $\tilde{\mathbf{X}}$, which results in a loss of information in the data structure. The approach uses functional data analysis to introduce a new framework for studying $\mathbf{X}$. Basis expansion techniques are employed to find analytic solutions for deformation variables such as rotation and parametrization deformations. Then, the generative model of $\mathbf{X}$ is investigated using a joint-principal component analysis approach. Numerical experiments on synthetic data and the *2dshapesstructures* datasets demonstrate how this new approach performs better at analyzing random planar curves than traditional functional data methods.

C1573: **Generalized functional probabilistic principal component analysis for longitudinal microbiome data analysis**
Presenter: **Xiangnan Xu**, Humboldt University of Berlin, Germany

Longitudinal microbiome studies are essential for understanding the dynamic microbial communities that inhabit various body sites and their interactions with host health, offering valuable insights for precision medicine. However, the analysis of longitudinal microbiome data presents challenges due to its high dimensionality, compositionality, overdispersion, and typically sparse sampling at irregular time points. Existing methods for dimension reduction, such as functional data analysis and tensor decomposition, often assume uniform sampling across individuals or focus on Gaussian-distributed variables. To address these, a novel generalized functional probabilistic principal component analysis (GFPPCA) framework is proposed that extends functional tensor decomposition to the setting of the exponential family distribution. GFPPCA integrates functional tensor decomposition with generalized probabilistic principal component analysis, enabling efficient and interpretable dimension reduction for complex functional microbiome data. An efficient alternating direction method of the multipliers-based algorithm is developed to estimate model parameters and validate GFPPCA on both simulated and real-world infant microbiome datasets. Results demonstrate that GFPPCA robustly and efficiently captures the primary information of the data across various parameter settings.

CC493 Room BH (SE) 2.12 RISK MANAGEMENT AND PORTFOLIO OPTIMIZATION Chair: Pavlo Mozharovskyi

C1339: **Optimal consumption and investment decisions with disastrous income risk**
Presenter: **Seyoung Park**, University of Nottingham, United Kingdom

An analytically tractable dynamic model of optimal consumption and investment decisions is developed with disastrous income risk in the context of Rietz's rare disaster risk hypothesis. The relations among consumption changes, aggregate income, disaster shock severity, and fiscal measures are first empirically explored in 55 countries during the COVID-19 period. Then, by empirical motivation, an important role of insurance is investigated with a focus on the recovery of income in a disaster. It highlights how the extent of disastrous income risk to which an agent is exposed and her income recovery post-disaster jointly affect the agent's optimal decisions. Overall, the availability of insurance can be particularly important for both the poor and the wealthy in the sense that they could even consume more, save less, and invest more post-disaster as long as their future income is (partly) recovered.

C1678: **Regime parity**
Presenter: **Florian Ielpo**, Centre Economie de la Sorbonne, France
Co-authors: Julien Royer, Selbi Muhammetgulyyeva

A new risk-based, long-run portfolio is introduced that bridges the gap between the multi-regime distribution of assets and risk-parity investing. Building on the seminal idea of the risk-parity portfolio, we relate regime-level information and portfolio allocation. When returns are influenced by latent regimes such as recession-expansion phases or inflation shocks, these non-stationarities can affect the unconditional distribution underlying the portfolio construction, leading the risk-parity allocation to load too much risk on the hedging asset. Our solution is the regime-parity portfolio, which is a linear combination of regime-specific risk-parity portfolios. We propose to weigh these regime-specific portfolios according to the steady-state probabilities of each regime, thus making use of the entire regime-level information. This new portfolio shows interesting features, notably better resistance to rare yet adverse regimes. We present the model and its salient features before showcasing different practical applications, highlighting the empirical interest of the approach.

C1500: **A bootstrap equality test to compare portfolio value at risk and expected shortfall**
Presenter: **Mario Maggi**, University of Pavia, Italy
Co-authors: Pierpaolo Uberti

In financial risk management, assessing the risk of an investment is fundamental, both from an absolute perspective and relative to some given benchmark. In practice, it often happens that an authority prescribes the risk measure to be used by the investors. For example, the international regulatory framework for banks Basel III imposes the expected shortfall at a given significance level in substitution to the value at risk. The aim is to propose a statistical test to measure if two investments are statistically significantly different in terms of value at risk and/or expected shortfall. The test is defined in non-parametric general settings. This permits comparing the risk of different investments without assuming a theoretical distribution of the returns. A method is also provided to build confidence intervals around the risk indicator of a given portfolio or asset. The only necessary assumption behind the methodology is that the risk measure verifies the translation invariance property. Although the focus is on the value at risk and the expected shortfall, the present approach is very general and can potentially be applied to a large class of risk measures. To support the proposal, some applications on real financial data are provided, and the results are discussed.

CP001 Room Auditorium POSTER SESSION Chair: Cristian Gatu

C0578: **On the robustness of random forests for genomic prediction and selection in breeding studies**
Presenter: **Vanda Lourenco**, NOVA University of Lisbon and NOVA.id.FCT, Portugal
Co-authors: Miguel Braga, Joao Lita da Silva

Real data analysis faces challenges due to potential violations of underlying model assumptions, such as errors or outliers. In linear regression, the presence of outliers can disrupt the normality assumption, leading to compromised parameter estimation and subsequent inferential results. Despite the effectiveness of machine learning methods like Random Forests (RF), susceptibility to data contamination remains a concern. The existing literature acknowledges the necessity for robust statistical techniques to address these issues, particularly in high-dimensional data analysis encompassing variable selection and prediction tasks. Enhancing the resilience of statistical methodologies is crucial for handling complex data scenarios and ensuring reliable analytical outcomes. While data contamination can manifest at both the response and covariate levels, this project primarily focuses on the former. The performance of the classical RF method is assessed via simulation while plugging in robust techniques to enhance its resilience against data contamination. Specifically, a synthetic animal dataset from the literature is employed, introducing various plausible contamination scenarios. The aim is to shed light on the implications of data contamination in genomic prediction and selection for breeding studies, offering insights into possible robust adaptations of RF that will help mitigate the challenges posed by certain types of contamination.

C1260: **Confidence intervals based on L-moments for population quantiles of the three-parameter kappa distribution**
Presenter: **Tereza Slamova**, Technical University of Liberec, Czech Republic

Conventional moments are traditionally used to describe features of a univariate probability distribution. An alternative approach uses quantities called L-moments, which are based on a linear combination of the expected value of the order statistics. Their main advantages in comparison to the conventional moments are the existence of all orders under only a finite mean assumption and higher robustness to outliers. L-moments also provide parameter estimators in a similar way as in the moment's method, and moreover, the method based on them is in some cases preferred over traditional methods for quantile estimation, specifically when such distributions with heavier tails than the normal distribution are fitted to smaller samples. A three-parameter kappa distribution is a special case of the four-parameter kappa distribution, and it has a heavier tail. The asymptotic confidence intervals for population parameters and quantiles of the three-parameter kappa distribution have been derived by using L-moments and compared to those obtained by conventional moments and maximum likelihood methods via Monte Carlo simulations.

C1298: **Duration analysis of Bitcoin trade with high-frequency transaction data**
Presenter: **Makoto Nakakita**, RIKEN, Japan
Co-authors: Teruo Nakatsuma

The aim is to understand the time series structure of duration between consecutive transactions of Bitcoin and identify its similarities and differences with other conventional financial assets such as stocks and commodities. For this purpose, a stochastic conditional duration (SCD) model is estimated using Bitcoin's high-frequency transaction data. To capture the effects of trade prices and volumes on the duration between transactions, those are incorporated into the SCD model as explanatory variables. Furthermore, the intraday seasonality of the duration is modeled with a Bernstein polynomial and simultaneously estimated with other parameters in the SCD model. The model estimation was performed using a Bayesian Markov chain Monte Carlo (MCMC) method. The estimation results suggested a positive relationship between the duration and the changes in trade prices but a negative relationship between the duration and the absolute values of the price changes, which are also known for the volatility in the stock market. The duration process of Bitcoin is also found to be strongly persistent, which is also found in other financial assets. In contrast, no clear pattern of intraday seasonality could be found in the Bitcoin market.

C1328: **Nonparametric estimation of reference curves**
Presenter: **Sandie Ferrigno**, INRIA Nancy and University Nancy Lorraine, France

In epidemiology, reference or standard curves are required to study fetal development in pregnancy. Values that lie outside the limits of these reference curves may indicate the presence of a disorder. Some classical empirical, parametric and semi-parametric methods, such as polynomial regression and LMS methods, are usually used to construct these curves. However, these classical methods build upon restrictive assumptions on estimated curves. The focus is on alternative nonparametric methods such as Nadaraya-Watson kernel estimation, local polynomial estimation, B-splines or cubic splines. The practical implementation of these methods to construct these curves requires working on smoothing parameters or choice of knots for the different types of nonparametric estimation. In particular, the optimal choice of these parameters is proposed. To fit these curves, we develop the R package quantCurves, an easy-to-use tool for practitioners, and a graphical interface to enable intuitive visualization of the results of the package.

C1430: **On statistical models and simulations for calculation of stockpile requirements for allergy-friendly foods**
Presenter: **Asanao Shimokawa**, Tokyo University of Science, Japan
Co-authors: Yuri Kominami

Calculating adequate food reserves for disasters is an important issue for local governments in many countries. For this reason, many countries have distributed standards and simulators for calculating food reserves with nutritional considerations. However, in many cases, consideration for the rapidly increasing number of patients with food allergies has not been addressed. To address this issue, statistical models are proposed to predict the number of food allergy cases and calculate reasonable food reserves based on information on the country's age-specific population composition. The proposed model uses a prediction interval that accounts for the prediction error in the number of food allergy patients. The properties of the proposed method are compared with those of existing methods under finite samples through simulations. As an example, the simulator is proposed, which developed to calculate the stockpile requirements of allergy-free foods in Japan that considering the expected number of food allergy sufferers.

C1432: **Prediction compensation-based recursive estimation with uniform quantization and random access protocol**
Presenter: **Raquel Caballero-Aguila**, Universidad de Jaen, Spain
Co-authors: Jiaxing Li, Jun Hu, Josefa Linares-Perez

The focus is on the recursive estimation problem for networked systems in the presence of uniform quantization and random access protocol with a prediction compensation strategy. The observations are quantized via the uniform quantizer to adapt to the digital transmission requirements. For the sake of improving communication efficiency and reducing resource consumption, the random access protocol is employed to schedule the obtained quantized observations, in which a prediction compensation strategy is utilized to mitigate the side effects of incomplete observations induced by the consideration of the random access protocol. Under the assumption that the evolution of the signal process is unknown and only the mathematical expectation and covariance of the signal process are given, a novel recursive least-squares linear estimation algorithm is proposed in light of an innovative approach. Finally, the validity of the developed recursive estimation algorithm is demonstrated by a numerical simulation experiment. In sum, the designed recursive estimation algorithm is beneficial for the development of signal processing in the context of networked environments, especially for cases including uniformly quantized observations and random access protocol.

C1446: **On confidence interval for correlation parameter in correlate gamma frailty models**
Presenter: **Kana Takamatsu**, Tokyo University of Science, Japan
Co-authors: Asanao Shimokawa

Survival analysis focuses on the time from a starting point, such as the initiation of surgery or treatment, until the occurrence of an event of interest. Many medical research datasets include information about subjects, such as age and sex, along with their survival times. One of the standard methods for analyzing the impact of these covariates on survival time is the Cox proportional hazards model. The genetic relationship between twins and families could potentially impact survival time. Therefore, a correlation frailty model with random effects is a method used to account for individual differences and non-independence within the same cluster. The distribution of random effects is often modeled using a gamma distribution. On the other hand, the properties of the estimated correlation parameter are not well-researched in many situations. The technique of constructing the confidence interval for the parameter is not well known. Therefore, the methods for estimating the confidence intervals of the correlation parameter in the correlated gamma frailty models are proposed using asymptotic distribution methods and bootstrap methods. These confidence intervals are compared through simulations and analysis of several examples.

C1448: **Functional mixed-effects models for function-valued traits in quantitative genetics**
Presenter: **Yilin Chen**, Kings College London, United Kingdom

Function-valued traits, such as growth trajectories, can be described as a function of a continuous index. These traits are assessed for continuous genetic variation using a mixed-effects model to decompose phenotypic variation into additive genetic and environmental components. A functional genetic mixed-effects model is proposed, using functional principal component analysis to derive a data-driven basis that improves the approximation of the underlying data structure. The model separates phase and amplitude variability by warping functions to address curve misalignment. It extends the lme4 package to model genetic data in a functional framework, with between-level correlations induced by post-multiplying the design matrix by the Cholesky factor of the additive genetic relationship matrix. Simulation studies show that the model captures the genetic covariance function effectively. Future focus will be on jointly analyzing warping functions and curve data.

C1450: **Comparison of the predictive values of two binary tests in the presence of categorical covariates**
Presenter: **Jose Antonio Roldan Nofuentes**, University of Granada, Spain

The comparison of the predictive values of two binary diagnostic tests is a topic of interest in the study of statistical methods for medical diagnosis and has been the subject of different studies in the statistics literature. In clinical practice, it is frequently found that when comparing diagnostic parameters, covariates are observed in all of the individuals in the sample, and it is necessary to adjust for these covariates. In this framework, a global hypothesis test is proposed to simultaneously compare the predictive values of two binary diagnostic tests when all of the individual's categorical covariates are observed. The global hypothesis test is solved through logistic regression and multinomial logit models. Simulation experiments were carried out to study the asymptotic behavior of the method proposed when a binary covariate is observed, and this was compared to the asymptotic behavior of the global test when the covariate is not used. In general terms, the method based on the regression models shows better asymptotic behavior than the other method. The results were applied to a real example.

C1721: **Generating time series for renewable energy analysis in power grids through bootstrap**
Presenter: **Andrea Marletta**, Universita degli Studi di Palermo, Italy
Co-authors: Giulia Marcon, Gianluca Sottile

As renewable energy becomes an increasingly important factor in electricity generation, accounting for the variability of its primary sources (such as wind speed and solar irradiance) is essential in energy grid design and power system analysis. In this context, simulating time series of renewable energy sources that reflect the characteristics of the original data helps understanding the impact of this variability on grid performances. A block bootstrap variant procedure is proposed to generate time series for wind speed, solar irradiance, or temperature with hourly frequency, addressing inherent seasonal trends. Data results as sufficiently representative, then simulated series exhibit similar behavior to the original input series. In order to check the adequacy of the proposed model, the evaluation of the procedure involves the main statistical properties such as mean, variance, and autocorrelation function which were compared between the original and simulated series. The approach allows simulation-based analysis of the impact of critical scenarios on different electric grids.

C1728: **Monte Carlo sampling benchmark suite**

Presenter: **Zeyu Ding**, TU Dortmund, Germany

A benchmark suite is introduced to assess the quality of Monte Carlo (MC) samples. Our suite enables quantitative comparisons using metrics like the sliced Wasserstein distance and other statistical measures. These are applied to both independent and identically distributed (iid) and correlated samples generated by MC methods, such as Markov Chain Monte Carlo or Nested Sampling. By gathering test statistics from repeated comparisons, we evaluate MC sampling quality. The suite includes diverse target functions of varying complexity and dimensionality, providing a flexible platform for testing sampling algorithms. Implemented as a Julia package, it allows users to select and extend test cases and metrics as needed. Users can run external sampling algorithms on these functions, input their samples, and obtain metrics that compare their quality to iid samples from our package. This standardized approach offers clear, quantitative measures of sampling quality, aiding researchers in validating and improving sampling methods.

Sunday 15.12.2024 13:40 - 15:20 Parallel Session H – CFECMStatistics2024

CI054 Room Auditorium ADVANCES IN MODELLING FINANCIAL AND ECONOMIC UNCERTAINTY Chair: Svetlana Makarova

C0159: **Shock persistence, uncertainty and news-driven business cycles**
Presenter: **Kalvinder Shields**, University of Melbourne, Australia
Co-authors: Kevin Lee, Guido Turnip

The purpose is to distinguish news about short-lived events from news about changes in longer-term prospects using surveys of expectations. Employing a multivariate GARCH-in-Mean model for the US illustrates how the different types of news influence business cycle dynamics. The influence of transitory output shocks can be relatively large on impact but gradually diminishes over two to three years. Permanent shocks drive the business cycle, generating immediate stock price reactions and gradually building output effects, although they have more immediate output effects during recessions through the uncertainties they create. Markedly different macroeconomic dynamics are found if these explicitly identified types of news or uncertainty feedbacks are omitted from the analysis.

C0160: **Unpacking economic uncertainty: Measuring the firm, sector and aggregate components**
Presenter: **Giulia Piccillo**, Maastricht University, Netherlands
Co-authors: Siavash Mohades, Tania Treibich

A novel method for measuring economic uncertainty at the firm, sector, and aggregate levels is introduced, using sales volatility, and it is validated by comparison with existing macroeconomic uncertainty measures. Compustat firms' data is used in the period 2000-2022 to construct the uncertainty measures for the U.S. economy. Findings highlight that 1) Macroeconomic conditions are the predominant source of firms uncertainty, 2) Diverse firm traits yield notable heterogeneity, and 3) The manufacturing sector exhibits the highest uncertainty among sectors. Findings shed light on the importance of firm and sectoral heterogeneity in studying uncertainty and its effects on economic activity.

C0161: **Unveiling the sources of economic uncertainty**
Presenter: **Andres Azqueta-Gavaldon**, Bank of Spain, Spain
Co-authors: Javier J Perez, Marina Diakonova, Corinna Ghirelli, Guillem Tobias

Detailed insights are provided into the nature and sources of economic uncertainty, offering a clearer and more precise understanding of the underlying factors driving uncertainty across different contexts. By focusing on the true sources of economic uncertainty, the aim is to harmonize the diverse indicators and metrics used in the field. By applying an established methodology across ten European countries using data in eight languages, the aim is to bring clarity and order to the myriad facets of economic uncertainty. A comprehensive new database is constructed to analyze and decompose news about economic uncertainty into various topics. Initially, topic modeling is used to identify the components of economic uncertainty. These components are then aggregated using the most natural and statistically validated techniques to ensure accurate grouping. This validation phase is crucial to confirm the robustness of the aggregated topics. In the second phase, these newly defined groupings are used to gain a deeper understanding of existing measures of uncertainty. Employing dimensional reduction techniques such as LASSO regression, the most significant components of uncertainty are identified, and their evolution is tracked over time. This approach also allows the examination of specific peaks in key uncertainty indicators like VIX, EPU, etc.

CO368 Room S-2.23 INDUSTRIAL STATISTICS (VIRTUAL) Chair: Wei-Heng Huang

C0357: **A semi-parametric CUSUM scheme using copula formulation for higher-order autocorrelated processes**
Presenter: **Yang-Li Liao**, Feng Chia University, Taiwan

In the Industry 4.0 era, rapid operations, characterized by the four Vs (volume, velocity, variety, and veracity), often encounter challenges related to autocorrelation in data streams. This manuscript presents a novel monitoring approach designed for autocorrelated processes, overcoming constraints on the order and magnitude of autocorrelations and distributional assumptions. The proposed phase II semi-parametric copula-based cumulative sum (CUSUM) control chart effectively detects shifts in location within continuous autocorrelated processes without relying on distributional assumptions. Average run length (ARL) performances against a distribution-free strategy demonstrate the proposed method's efficacy, showcasing adaptability in handling various autocorrelation magnitudes. Application of this CUSUM method in the selective laser melting (SLM) process produces prompt signals, validating its effectiveness in real-world datasets.

C0362: **Exact average coverage probabilities and confidence coefficients of intervals for a difference of success probabilities**
Presenter: **Chung-Han Lee**, National Chung Cheng University, Taiwan
Co-authors: Yu-Hsuan Tai

For a confidence interval of a parameter in two independent binomial distributions, the coverage probability is a bivariate variable function of parameters, and the confidence coefficient is the infimum of the coverage probabilities. In general, the exact confidence coefficients and average coverage probabilities of intervals for differences of binomial proportions are unknown, and they are usually approximated by taking the minimum and mean of the coverage probabilities at randomly chosen points in the parameter space. The aim is to propose methodologies for calculating the exact confidence coefficients and average coverage probabilities of intervals for a risk difference, respectively. Therefore, the proposed methods are used to illustrate the performance of existing intervals and make recommendations.

C0895: **A generalized distribution-free multivariate control chart with improved post-signal diagnostics**
Presenter: **Chase Holcombe**, University of South Alabama, United States

Multivariate statistical process control (MSPC) charts are particularly useful when the need arises to simultaneously monitor several quality characteristics of a process. Most control charts in MSPC assume that the quality characteristics follow some parametric multivariate distribution. Distribution-free MSPC charts are attractive, as they can guarantee in-control (IC) or null performance of the control chart without the assumption of a parametric multivariate process distribution. Utilizing an existing tolerance interval, a simple phase II Shewhart-type distribution-free MSPC chart is constructed and proposed for individual and subgrouped observations. The proposed charting methodology preserves the original scale of measurements and can easily identify out-of-control (OOC) variables after a signal, which are both important practical advantages in the multivariate setting. The proposed control chart is attractive as it is easy to construct, visualize, and interpret, is exactly distribution-free, requires no complex parameter estimation calculations for implementation, comes with a natural and simple post-signal diagnostic mechanism, and only requires a modestly large reference sample size for small to moderate dimensions. IC and OOC performance is discussed. Effects of contamination and high dimensionality are also briefly discussed.

C0602: **The performance of S control charts for the lognormal distribution with estimated parameters**
Presenter: **Wei-Heng Huang**, National Taipei University, Taiwan

Control charts, one of the powerful tools in statistical process control (SPC), are widely used to monitor and detect out-of-control processes in the manufacturing industry. Many researchers have pointed out the effects of using estimated parameters on the average run length (ARL) performance metric. Most of the previous literature has studied the expected value of the average run length (AARL) and the standard deviation of the average run length (SDARL) to evaluate the performance of control charts. The purpose is to study the performance of three S control charts, the Shewhart S-chart, the median absolute deviation (MAD) control chart, and the lognormal S control chart, for a lognormal distribution in terms of the AARL and SDARL. Simulation results indicate the sample size that will reach the specified control ARL value is very large. The lognormal S control chart has a smaller SDARL value than the other two S-charts. A real example is used to demonstrate how the proposed chart can be applied in practice.

CO072 Room S-2.25 SPORTS ANALYTICS **Chair: Christophe Ley**

C0779: **Injury prediction in soccer with conventional statistical approaches and machine learning models**
Presenter: **Ina-Marie Berendes**, TU Dortmund University, Germany
Co-authors: Alexander Gerharz, Andreas Groll, Mathias Kolodziej

Injuries in professional soccer happen frequently and are problematic for players and their clubs. The prevention of injuries can be addressed from a statistical perspective to identify connections. One possibility is modeling the binary injury status of players. The approaches for injury modeling in young professional soccer players compared here are Lasso regularized logistic regression, naive Bayes, linear discriminant analysis, k-nearest neighbors, classification trees, random forests, XGBoost, and support vector machines. They are employed to predict the injury probability and status of players. A parallel cross-validated procedure and several quality measures are used to compare the different methods. A post-Lasso logistic regression model with a decreased penalty emerges as the overall best model with a sensitivity of 0.773, a specificity of 0.529, an AUC of 0.672, an accuracy of 0.625, a predictive likelihood of 0.593, and a Brier score of 0.228. It contains three features relevant for injury prediction: the players' postural control sway under static conditions, the concentric knee extension torque, and the transversal plane moment of the hip during a single-leg drop landing task. An XGBoost model reaches a slightly higher accuracy of 0.661 but doesn't match the Lasso models performance regarding the other measures.

C0860: **A bivariate extension of the regularized adjusted plus-minus model**
Presenter: **Luca Grassetti**, University of Udine, Italy
Co-authors: Valentina Mameli, Michele Lambardi di San Miniato

From the game's results point of view, the predictive power of model-based player performance measures, such as regularized adjusted plus-minus (RAPM), is typically poor. To enhance this feature, one can consider a bivariate model formulation like in the double Poisson specification, typical of soccer game results prediction. First, the home and away team scores are computed separately, and each score is specified by considering the effects of the combination of offensive and defensive non-negative players. Second, the scores can be computed over equally spaced periods. This solution implies that, for each period, the presence of players on the field must be measured by considering their usage percentage. The resulting model specification is equivalent to the model based on dummies, but it eases the bivariate generalization of the model from a computational point of view. The 2022-2023 NBA play-by-play data are analyzed to evaluate the proposal. The results of the empirical analyses show that bivariate RAPM models inherit the advantages of model-based procedures from the player's performance point of view. Moreover, the novel model specification can be used to describe the game's progress and try to predict the results of game periods.

C0962: **Comparison of prediction models for survival analysis of running-related injuries**
Presenter: **Katarzyna Szczerba**, University of Luxembourg, Luxembourg
Co-authors: Christophe Ley, Laurent Malisoux, Daniel Theisen

As awareness grows about the benefits of physical activities, there is an increasing need for advanced tools in sports medicine. Running, while popular and beneficial, has a high injury rate, leading to severe long-term impacts. This raises a crucial question: what are the primary risk factors for running injuries, and what preventive measures should be considered? To address this issue, the Luxembourg Institute of Health conducted a randomized controlled trial in 2017 involving 848 runners, with a 6-month follow-up. While several insightful papers resulted, none used machine learning due to its 'black box' nature and initial underperformance compared to traditional Cox models. The aim is to develop a superior machine learning model to identify risk factors for running-related injuries using innovative interpretable machine learning (IML) methods. Using 10-fold cross-validation with the concordance index as the evaluation metric, we found that a gradient-boosted Cox proportional hazards model with regression trees as base learners outperformed all other models. To ensure explainability, SHapley Additive exPlanations (SHAP) is also employed, and additional statistical analyses are conducted. This case study illustrates that sports scientists can achieve deeper insights into their data by employing advanced machine learning models with interpretable machine learning (IML) methods.

C1059: **From data to dominance: The new era of sports analytics**
Presenter: **Lorena Martin**, University of Southern California, United States

Sports analytics is at the cusp of innovation integration in the competitive landscape, offering unprecedented insights into athlete performance and health. The integration of cutting-edge technologies, including force plates, motion analysis, and wearable devices, along with physiological and psychological AI tools, is explored to create a comprehensive framework for optimizing athletic performance and preventing injuries. By harnessing this data, teams can make unparalleled decisions that enhance training regimens, game strategies, and overall player well-being. The rapid development and deployment of innovative AI tools have brought about a significant transformation in professional sports. Teams and players are increasingly adopting these unique technologies to gain a competitive edge, while some stakeholders express concerns about data privacy, accuracy, and the potential for over-reliance on technology. These challenges are addressed, and the varied responses are discussed within the sports industry, highlighting both the excitement and resistance from different sectors. "From Data to Dominance" showcases the exclusive potential of sports analytics.

CO063 Room S-1.01 HITEC: TOPICS ON HIGH DIMENSIONAL STATISTICS **Chair: Andreas Artemiou**

C0203: **Kernel association rotation analysis: A kernel-based projection of continuous data**
Presenter: **Kimon Ntotsis**, University of Leicester, United Kingdom
Co-authors: Andreas Artemiou, Alexandros Karagrigoriou

In high-dimensional data analysis, datasets often contain both linear and non-linear associations. Traditional dimensionality reduction techniques, like principal component analysis, focus primarily on linear associations, potentially missing important non-linear connections. This gap necessitates new methodologies capable of capturing the full spectrum of relationships. The aim is to present the kernel association rotation analysis (KARA), an unsupervised approach that excels in projecting the original data into a lower-dimensional space while retaining the majority of the original data's variance. KARA employs the kernel association coefficient (KAC), a measure designed to capture both linear and non-linear associations with high accuracy. Using a standardized 2-degree polynomial kernel function, KARA provides a comprehensive evaluation of feature relationships and data projection into a lower-dimensional space. Through simulations and real-world case studies, KARA's performance is evaluated and compared to other techniques using Gaussian process regression.

C0232: **Inverse regression with column selection: A unified generalization of inverse regression via adaptive column selection**
Presenter: **Yin Jin**, Zhejiang University, China

A bottleneck of sufficient dimension reduction (SDR) in the modern era is that, among numerous methods, only the sliced inverse regression (SIR) is generally applicable under high-dimensional settings. The higher-order inverse regression methods, which form a major family of SDR methods that are superior to SIR at the population level, suffer from the dimensionality of their intermediate matrix-valued parameters that have an excessive number of columns. The generic idea of using a small subset of columns of the matrix-valued parameter for SDR estimation is proposed, which breaks the convention of using the ambient matrix for the higher-order inverse regression methods. With the aid of a quick column selection procedure, these methods, as well as their ensembles towards sparsity, are then generalized under the high-dimensional settings in a uniform manner that resembles sparse SIR and without additional assumptions. This is the first promising attempt in the literature to free the higher-order inverse regression methods from their dimensionality, which facilitates the applicability of SDR. The gain of column selection with respect to SDR estimation efficiency is also studied under the fixed-dimensional settings. Simulation studies and a real data example are provided at the end.

C0501: **Sufficient dimension reduction for conditional quantiles for functional data**
Presenter: **Eliana Christou**, University of North Carolina at Charlotte, United States
Co-authors: Eftychia Solea, Shanshan Wang, Jun Song

Functional data analysis is an important research area with the potential to transform numerous fields. However, existing work predominantly relies on the more traditional mean regression methods, with surprisingly limited research focusing on quantile regression. Furthermore, the infinite-dimensional nature of the functional predictors necessitates the use of dimension-reduction techniques. Therefore, this gap is addressed by developing dimension-reduction techniques for the conditional quantiles of functional data. The convergence rates are derived from the proposed estimators, and their finite sample performance is demonstrated using simulation examples and a real dataset from fMRI studies.

C0887: **Multi-response linear regression estimation based on low-rank pre-smoothing**
Presenter: **Sandipan Roy**, University of Bath, United Kingdom
Co-authors: Matthew Nunes, Xinle Tian, Alex Gibberd

Pre-smoothing is a technique aimed at increasing the signal-to-noise ratio in data to improve subsequent estimation and model selection in regression problems. However, pre-smoothing has thus far been limited to the univariate response regression setting. Motivated by the widespread interest in multi-response regression analysis in many scientific applications, a technique for data pre-smoothing is proposed in this setting based on low-rank approximation. Theoretical results are established on the performance of the proposed methodology and quantify its benefit empirically in a number of simulated experiments. The proposed low-rank pre-smoothing technique is also demonstrated on real data arising from the environmental sciences.

CO149 Room S-1.04 COMPUTATIONAL ASYMPTOTIC STATISTICS FOR STOCHASTIC PROCESSES Chair: Nakahiro Yoshida

C0422: **High-frequency market manipulation detection with a Markov-modulated Hawkes process**
Presenter: **Ioane Muni Toke**, CentraleSupelec, France
Co-authors: Timothee Fabre

The focus is on a self-exciting point process defined by a Hawkes-like intensity and a switching mechanism based on a hidden Markov chain. Previous works in such a setting assume constant intensities between consecutive events. The model is extended to general Hawkes excitation kernels that are piecewise constant between events. An expectation-maximization algorithm is developed for the statistical inference of the Hawkes intensities parameters as well as the state transition probabilities. The numerical convergence of the estimators is extensively tested on simulated data. The model is applied to high-frequency cryptocurrency market data on a top centralized exchange. The focus is on two high-frequency market manipulation strategies which are wash trading using many orders and pinging. The goodness of fit of the model is benchmarked with the Markov-modulated Poisson process, and the effectiveness of the model is demonstrated in detecting past suspicious activities over a large out-of-sample dataset.

C1247: **Rough volatility estimation in a fully general framework**
Presenter: **Mathieu Rosenbaum**, Ecole Polytechnique, France

Rough volatility models are nowadays very popular among practitioners due to their remarkable ability to fit with great accuracy market data with a very small number of parameters. The issue of estimating with precision the rough nature of volatility in a very general setting is addressed, including jumps and microstructure effects. Optimal procedures are designed, providing speed of convergence and limit theorems. The results are also illustrated on real data.

C0999: **Drift estimation for rough diffusion models under a small noise asymptotic assumption**
Presenter: **Arnaud Gloter**, Universite d Evry Val d Essonne, France
Co-authors: Nakahiro Yoshida

A stochastic process X solution of a rough Volterra equation driven by a semimartingale process Z is considered. It is assumed that the diffusion coefficient of Z is proportional to some constant ε tending to zero. The focus is on the estimation of the drift parameter from continuous and discrete observations of the process X on a compact time interval $[0,T]$. The asymptotic behavior of two kinds of estimators is studied, one based on a trajectory fitting method and one obtained by an approximate likelihood method.

C0577: **Analytical approximations for diffusion processes with asymptotic expansion**
Presenter: **Emanuele Guidotti**, University of Lugano, Switzerland
Co-authors: Nakahiro Yoshida

Diffusion processes are a class of models that plays a prominent role in describing the time-continuous evolution of phenomena in the natural and social sciences. However, only in very few cases can the stochastic differential equation driving the process be analytically solved. Based on the perturbation method, asymptotic expansion formulas are presented to generate accurate approximations of the solution of arbitrary diffusions and describe their implementation.

CO183 Room S-1.06 NEW CHALLENGES IN MODERN STATISTICS AND DATA ANALYSIS (VIRTUAL) Chair: Yichuan Zhao

C0885: **Causal inference for time-to-event data with a cured subpopulation**
Presenter: **Yi Wang**, Shanghai University of International Business and Economics, China

When studying the treatment effect on time-to-event outcomes, it is common that some individuals never experience failure events, which suggests that they have been cured. However, the cure status may not be observed due to censoring, which makes it challenging to define treatment effects. Current methods mainly focus on estimating model parameters in various cure models, ultimately leading to a lack of causal interpretations. To address this issue, two causal estimands are proposed, the timewise risk difference and mean survival time difference, in the always-uncured based on principal stratification as a complement to the treatment effect on cure rates. These estimands allow the study of the treatment effects on failure times in the always-uncured subpopulation. The identifiability is shown using a substitutional variable for the potential cure status under the ignorable treatment assignment mechanism; these two estimands are identifiable. Estimation methods are also provided using mixture cure models. The approach is applied to an observational study that compared the leukemia-free survival rates of different transplantation types to cure acute lymphoblastic leukemia. The proposed approach yielded insightful results that can be used to inform future treatment decisions.

C1244: **Identification of causal effects via instrumental variables from an auxiliary heterogeneous population**
Presenter: **Wei Li**, Renmin University of China, China

Evaluating causal effects in a primary population of interest with unmeasured confounders is challenging. Although instrumental variables (IVs) are widely used to address unmeasured confounding, they may not always be available in the primary population. Fortunately, IVs might have been used in previous observational studies on similar causal problems, and these auxiliary studies can be useful to infer causal effects in the primary population, even if they represent different populations. However, existing methods often assume homogeneity or equality of conditional average treatment effects between the primary and auxiliary populations, which may be limited in practice. The aim is to remove the homogeneity requirement and establish a novel identifiability result, allowing for different conditional average treatment effects across populations. A multiply robust estimator that remains consistent despite partial misspecifications of the observed data model is also constructed and achieves local efficiency if all nuisance models are correct. The proposed approach is illustrated through simulation studies. The approach is finally applied by leveraging data from lower-income individuals with cigarette prices as a valid IV to evaluate the causal effect of smoking on physical functional status in higher-income groups where strong IVs are not available.

C1530: **Tackling the efficiency paradox for data fusion with many external studies**

Presenter: **Jingyue Huang**, University of Pennsylvania, United States

The problem of integrating individual data is considered from an internal study, and many summary statistics are derived from a separate external study. The research uncovers a paradox: using multiple external summary statistics could worsen the finite-sample performance of data fusion methods, even when the statistics are unbiased and have low variability. By introducing a linear regression representation for data-fused estimators, this paradox is characterized by an inherent trade-off that fewer external studies for integration yield smaller estimation variance but result in a larger discrepancy relative to the semiparametric efficiency bound. A lasso-type regularization method is further proposed to balance the trade-off. The theoretical analysis shows that the semiparametric efficiency bound remains achievable if the number of informative external studies does not grow too quickly. The applicability of the method is also demonstrated to federated transfer learning with structural missingness, which may be of independent interest. The effectiveness of the proposed method is evidenced by simulations and a real-world study.

C0434: **Smoothed empirical likelihood for the difference of two quantiles with the paired sample**

Presenter: **Yichuan Zhao**, Georgia State University, United States

Co-authors: Pangpang Liu

A novel smoothed empirical likelihood method is proposed for the difference of quantiles with paired samples. While the empirical likelihood for the difference of two quantiles with independent samples has been studied, it is crucial to develop a statistical procedure that accounts for the dependence between paired samples. To this end, two estimating equations are proposed for the difference of two quantiles, and a nuisance parameter is introduced in the smoothed empirical likelihood framework. It is demonstrated that the approach yields a limiting distribution that follows Wilks' theorem. Extensive simulation studies confirm that the smoothed empirical likelihood method outperforms the normal approximation and method M in most cases. Finally, the usefulness of the proposed method is illustrated by applying it to a real-world data set, estimating the interval of the quantile difference of GDP between different years.

CO221 Room S-1.27 ALL MODELS ARE WRONG BUT MANY ARE USEFUL Chair: Lucas Mentch

C1622: **Forward stability and model path selection**

Presenter: **Lucas Mentch**, University of Pittsburgh, United States

Most scientific publications follow the familiar recipe of (i) obtaining data, (ii) fitting a model, and (iii) commenting on the scientific relevance of the effects of particular covariates in that model. This approach, however, ignores the fact that there may exist a multitude of similarly-accurate models in which the implied effects of individual covariates may be vastly different. This problem of finding an entire collection of plausible models has also received relatively little attention in the statistics community, with nearly all of the proposed methodologies being narrowly tailored to a particular model class and/or requiring an exhaustive search over all possible models, making them largely infeasible in the current big data era. The idea of forward stability is developed, and a novel, computationally-efficient approach is proposed to finding collections of accurate models referred to as model path selection (MPS). MPS builds up a plausible model collection via a forward selection approach and is entirely agnostic to the model class and loss function employed. The resulting model collection can be displayed in a simple and intuitive graphical fashion, easily allowing practitioners to visualize whether some covariates can be swapped for others with minimal loss.

C0179: **On the existence of simpler-yet-accurate models**

Presenter: **Lesia Semenova**, Microsoft Research, United States

Finding optimal, sparse, accurate models of various forms (such as linear models with integer coefficients, rule lists, and decision trees) is generally NP-hard. Often, it is unknown whether the search for a simpler model will be worthwhile, and thus the effort to find one is not undertaken. The purpose is to address an important practical question: For which types of datasets interpretable models are expected to perform as well as black-box models? A mechanism of the data generation process is presented, coupled with choices usually made by the analyst during the learning process, that leads to the existence of simpler-yet-accurate models. This mechanism indicates that such models exist in practice more often than one might expect.

C0467: **Confidence sets for high-dimensional variable selection**

Presenter: **Davide Ferrari**, Free University of Bozen/Bolzano, Italy

The aim is to present a methodology for constructing variable selection confidence sets for high-dimensional regression models. Unlike traditional variable selection approaches, the new methodology goes beyond relying on a singular model obtained through a specific model selection criterion. Instead, it constructs a set of regression models encompassing the true model with a user-specified level of confidence. In noisy high-dimensional datasets, discriminating competing models is more challenging and generally results in broader confidence sets; conversely, in informative datasets, the confidence set contains relatively few valuable models. Within the confidence set methodology, the characteristics of lower boundary models are examined, defined as the set of most parsimonious models included in the confidence sets. Leveraging insights from the confidence set and lower boundary models, natural measures are explored to characterize overall model uncertainty and the importance of individual variables.

C0181: **Model class selection**

Presenter: **Ryan Cecil**, University of Pittsburgh, United States

Co-authors: Lucas Mentch

Classical model selection generally seeks to find a single model within a particular class that optimizes some pre-specified criteria, such as maximizing a likelihood or minimizing a risk. More recently, there has been an increased research focus on model set selection (MSS), where the aim is to identify a subset of all models such that no model in the selected subset is significantly worse than the empirically optimal one. This subset of optimal models is sometimes referred to as the Rashomon set. The MSS framework is generalized one step further by introducing the idea of model class selection (MCS). In MCS, multiple model collections or classes are assumed, and all such collections that do (or do not) contain at least one optimal model are sought for identification. As a direct consequence, for particular datasets we are able to investigate formally whether classes of simpler and more interpretable statistical models are able to perform on par with more complex black-box machine learning models. In other words, as it has become relatively common for practitioners to rule out classical models a priori because the data is assumed to be too large and/or complex, another means of evaluating whether (and when) such actions are justified is proposed. A variety of simulated and real-data experiments are provided.

CO135 Room K0.16 RECENT ADVANCES IN NETWORKS AND HIGH DIMENSIONAL DATA Chair: Vince Lyzinski

C0509: **Augmented degree correction for bipartite networks with applications to recommender systems**

Presenter: **Benjamin Leinwand**, Stevens Institute of Technology, United States

Co-authors: Vladas Pipiras

In recommender systems, users rate items and are subsequently served other product recommendations based on these ratings. Even though users usually rate a tiny percentage of the available items, the system tries to estimate unobserved preferences by finding similarities across users and across items. The observed rating data is treated as partially observed, dense, weighted, bipartite networks. For a class of systems without outside information, an approach developed for dense, weighted networks is adapted to account for unobserved edges and the bipartite nature of the problem. The approach begins with clustering both users and items into communities and locally estimates the patterns of ratings within each subnetwork induced by restricting attention to one community of users and one community of items community. The local fitting procedure relies on estimating local sociability parameters for every user and item and selecting the function to determine the degree correction contours that best model the underlying data. On a joke ratings data set, the proposed model performs better than existing alternatives in relatively sparse settings, though other approaches achieve better results when more data is available. The results indicate that despite struggling to pick up subtler signals, the proposed approach to recovery of large-scale, coarse patterns may still be useful in practical settings where high sparsity is typical.

C0795: **Signed diverse multiplex networks: Clustering and inference**
Presenter: **Marianna Pensky**, University of Central Florida, United States

The purpose is to introduce a signed generalized random dot product graph (SGRDPG) model, which is a variant of the generalized random dot product graph (GRDPG), where, in addition, edges can be positive or negative. The setting is extended to a multiplex version, where all layers have the same collection of nodes and follow the SGRDPG. The only common feature of the layers of the network is that they can be partitioned into groups with common subspace structures, while otherwise, matrices of connection probabilities can be all different. The setting above is extremely flexible and includes a variety of existing multiplex network models as its particular cases. By employing novel methodologies, strongly consistent clustering of layers and high accuracy of subspace estimation is ensured, which were attained previously only in much simpler models. All algorithms and theoretical results remain true for both signed and binary networks. In addition, it shows that keeping signs of the edges in the process of network construction leads to a better precision of estimation and clustering and, hence, is beneficial for tackling real-world problems such as, for example, analysis of brain networks.

C0806: **Multi-source matrix data integration via embedding alignment**
Presenter: **Runbing Zheng**, Johns Hopkins University, United States
Co-authors: Minh Tang

Motivated by the increasing demand for multi-source data integration in various scientific fields, the problem of matrix completion is studied where the observable information of the whole to-recover matrix presents certain block-wise missing structures. The interest is in recovering an underlying low-rank and large-scale matrix, of which can only be observed several noisy submatrices of certain parts of entities. An algorithm is proposed to explicitly integrate all information revealed in the noisy submatrices, thereby efficiently estimating the underlying truth. Specifically, the proposed algorithm first estimates entity embeddings for each observed submatrix. It then aligns the embeddings between submatrices of overlapping entities and finally aggregates the aligned embeddings over all submatrices to recover the whole large matrix of interest. The asymptotic analysis showcases that the algorithm can entrywisely recover the underlying truth, and moreover, the entrywise fluctuations of the estimate are proven to be mean-zero normally distributed. The simulation and real data studies show that the algorithm is efficient and effective for this structured matrix completion problem.

C0832: **Matching and mixing: Matchability of graphs under Markovian error**
Presenter: **Zhirui Li**, University of Maryland College Park, United States
Co-authors: Keith Levin, Zhiang Zhao, Vince Lyzinski

Theoretical results on the matchability of graphs are presented under dependent Markovian noise. The theories are built upon a lamplighter-like model, which, unlike most existing work in the graph-matching literature, does not assume the independence of noise on the edges. By comparing the Markovian mechanism of the lamplighter walk and the ability to de-anonymize (i.e., match) the noisy network, it is shown that under classical homogeneous Erdos-Renyi models, the anonymization time and the mixing time of the random walk are both of order $O(n^2\text{polylog}(n))$. It is further demonstrated that for more structured models (e.g., the stochastic block model), anonymization can occur in $O(n^{\alpha}\text{polylog}(n))$ time for $\alpha < 2$, indicating that matching fails before the Markovian noise globally mixes. These established time bounds are demonstrated using both simulated graph models in the settings of Erdos-Renyi and SBM graphs, as well as real-world datasets such as the Facebook friendship network and the European research institution email communication network.

CO096 Room K0.18 INNOVATIVE STATISTICAL APPROACHES FOR CLIMATE CHANGE STUDIES Chair: Andriette Bekker

C0695: **An axial regression model to detect vegetation growth patterns**
Presenter: **Marco Mingione**, Roma Tre, Italy

A mixture of axial regression models is proposed for bivariate observations of axial and circular variables. It is noted that an axial variable is defined on the semicircle due to the lack of information about the direction of the propagation. Therefore, the data lies on the Cartesian product of a circle and a semicircle. A valid density function is defined with periodic behavior on this complex manifold by exploiting the bivariate wrapped Cauchy density to induce dependence between the two angular measurements. The proposed parametrization allows for the derivation of conditional distributions in closed form, and straightforward regression models are built. Estimation is obtained via maximum likelihood by exploiting the EM algorithm. The model's performances are illustrated on an original dataset of stripe and wind directions recorded on Marion Island (South Africa).

C0953: **Exploring pollutant patterns through graphical elastic net**
Presenter: **Priyanka Nagar**, Stellenbosch University, South Africa
Co-authors: Azam Kheyri, Andriette Bekker

An analysis of the global network of an air quality metric is presented using data procured from various urban centers. By leveraging advanced network analysis techniques, a novel framework is proposed to decipher the intricate dynamics of a pollutant across various locations. Distinctly rooted in graph theory, the approach deviates from traditional autoregressive models by integrating a simultaneous transmission component. This integration enables a more holistic understanding of the pollutant interactions across different sites. Empirical findings provide insight into the complex global network of the pollutant, identifying key points and clarifying its linkages.

C1099: **INLA: A computational tool for climate modeling**
Presenter: **Janet Van Niekerk**, King Abdullah University of Science and Technology, Saudi Arabia
Co-authors: Haavard Rue

Climate modeling often involves large data and/or complex models. In this realm, Bayesian modeling offers great promise to explain complex dependencies, but the computational complexity involved can deter the implementation of such models. Recent advances in the INLA methodology employ a Variational Bayes correction instead of the nested Laplace approximations, which results in INLA performing inference even faster and for more complex models or huge data. The purpose is to briefly introduce the modern INLA methodology and illustrate some examples of global climate models.

C1110: **Detecting anomalies in emission trends: A statistical approach for environmental policy in Portugal**
Presenter: **Ana Borges**, CIICESI, ESTG, Politacnico do Porto, Portugal

Co-authors: Clara Cordeiro, M Rosario Ramos, Mariana Carvalho

To improve understanding of variations in emissions that serve as key indicators of environmental quality and the effectiveness of policies, a statistical methodology is proposed for detecting anomalies in the time series dataset of monthly emissions in Portugal. The focuse is on CO, NH3, NMVOC, NOx, OC, PM2.5, PM10, and SO2 emissions from 2000 to 2022. This information on changes is crucial so that policies may be designed for effective strategies toward emission reduction and improvement of air quality. The analytical method adapts a technique that has been developed for analysing hydrological-related time series data. First, the seasonal trend decomposition based on the Loess method is used to decompose the time series. Thereafter, a breakpoint analysis is undertaken to search for drastic changes in the emission trends for the seasonally adjusted data. Mann-Kendall test and Sen's slope estimator are used to discover the presence of significant increases or decreases in emissions. The application of this methodology to emissions data has been successful in the identification of breakpoints corresponding to significant changes in emission patterns. This is highly useful in environmental surveillance since it facilitates proactive measures to abate air pollution and its impact on health and the environment.

CO091 Room K0.20 RECENT DEVELOPMENTS ON DATA DEPTH AND FUNCTIONAL DATA ANALYSIS Chair: Sara Lopez Pintado

C0810: **Regularized halfspace depth: Practical insights and applications**
Presenter: **Hyemin Yeon**, Kent State University, United States
Co-authors: Xiongtao Dai, Sara Lopez Pintado

Data depth is a powerful nonparametric tool originally proposed to rank multivariate data from the centre outward. In this context, one of the most archetypical depth notions is Tukey's halfspace depth. In the last few decades, notions of depth have also been proposed for functional data. However, Tukey's depth cannot be extended to handle functional data because of its degeneracy. A new halfspace depth for functional data is proposed, which avoids degeneracy by regularization. The halfspace projection directions are constrained to have a small reproducing kernel Hilbert space norm. Desirable theoretical properties of the proposed depth, such as isometry invariance, maximality at center, monotonicity relative to the deepest point, upper semi-continuity, and consistency, are established. Moreover, depending on the regularization, the regularized halfspace depth can rank functional data with varying emphasis in shape or magnitude. A new outlier detection approach is also proposed, capable of detecting shape and magnitude outliers. It is applicable to trajectories in L2, a very general space of functions that include non-smooth trajectories. Based on extensive numerical studies, the methods are shown to perform well in detecting different types of outliers. Three real data examples showcase the proposed depth notion. We focus on the practical aspects of the proposed regularized halfspace depth.

C0833: **Distributionally robust halfspace depth**
Presenter: **Pavlo Mozharovskyi**, LTCI, Telecom Paris, Institut Polytechnique de Paris, France
Co-authors: Jevgenijs Ivanovs

Statistical data depth function measures the centrality of an observation concerning a distribution or a data set by a value between 0 and 1 while satisfying certain postulates regarding invariance, monotonicity, and convexity. It constitutes a contemporary domain of rapid development that meets growing demand in industry, economy, social sciences, etc. Being one of the most studied depth notions, Tukey's halfspace depth can be seen as a stochastic program, and as such, it suffers from the optimizer's curse: a limited training sample results in a poor out-of-sample performance. A generalized halfspace depth concept is proposed, relying on the recent advances in distributionally robust optimization, where halfspaces are examined using the worst-case distribution in the Wasserstein ball around the empirical law. This new depth can be seen as a smoothed and regularized classical halfspace depth retrieved as the radius of the Wasserstein ball vanishes. It inherits the latter's main properties and enjoys various new attractive features, such as continuity and strict positivity beyond the convex hull of the support. Numerical illustrations of the new depth and its advantages are provided, and some fundamental theory is developed. In particular, the upper-level sets and the median region are studied, including their breakdown properties, and distributionally robust halfspace depth is applied to the tasks of outlier detection and supervised classification.

C1015: **Robust depth-based registration for multivariate functional data**
Presenter: **Ana Arribas-Gil**, Universidad Carlos III de Madrid, Spain
Co-authors: Sara Lopez Pintado

In the context of multivariate functional data with individual phase variation, a robust depth-based approach is presented to estimate the main pattern function in specific time-warping models. In particular, the latent-deformation model is considered, in which the different components of a multivariate functional variable are also time-distorted versions of a common template function. Rather than focusing on a particular functional depth measure, the necessary conditions of a depth function are discussed to provide a consistent estimation of the central pattern, considering different model assumptions. The method's performance and robustness are evaluated against atypical observations and violations of the model assumptions through simulation and illustrate its use on two real data sets.

C0972: **The quantile integrated depth with applications to noisy functional data**
Presenter: **Sara Lopez Pintado**, Northeastern University, United States
Co-authors: Stanislav Nagy, Todd Ogden, Man Luo

Functional data analysis involves data for which the basic unit of observation is a function or image. The development of robust exploratory tools and inferential methods is very much needed since few assumptions can be made about the generating process. Data depth, a well-known nonparametric tool for analyzing functional data, provides a rigorous method for ranking a sample of curves from the center outwards, allowing for robust inference and outlier detection. Several notions of depth for functional data have been introduced in the last few decades. A new family of depths is developed, termed quantile integrated depth (QID), that is based on integrating up to the K-th quantile of the univariate depths. It is shown that this new family of depths has desirable properties, including a type of invariance, maximality at the center, and monotonicity with respect to the deepest point. In addition, since functional data are commonly observed with noise, we explore the effect of noise on different notions of depth. Compared to alternatives, the proposed QID is shown to be robust and perform well on noisy functional data. Procedures are also discussed for choosing the optimal tuning parameter K for QID.

CO157 Room K0.50 RECENT ADVANCES IN DESIGN AND ANALYSIS OF EXPERIMENTS Chair: Chenlu Shi

C0387: **Relaxed greedy packing for nested space-filling designs**
Presenter: **Luc Pronzato**, CNRS - Universite Cote d'Azur, France

Various relaxations of the greedy packing algorithm are considered for the construction of nested designs on a given compact set X, with a guaranteed lower bound on packing and covering efficiency for each design size. Relaxation can include randomness. Although the stochastic properties of the nested random designs that are generated are much more difficult to study than those of designs generated by the now popular determinant point processes, (i) the construction of a design $\{x_1,\ldots,x_n\}$ of arbitrary size n is straightforward, (ii) guaranteed packing and covering performance is available for all sub-designs $\{x_1,\ldots,x_k\}$, $k \leq n$. When X is the hypercube $[0,1]^d$, projections onto canonical subspaces can be taken into account, random Latin hypercubes being special cases. Relaxation can correspond to greedy minimization of the energy for a positive definite kernel K, which defines a correlation function for a random process on X, with Matérn kernels as special cases. It is shown that when the correlation length tends to zero fast enough, the sequence of nested designs is asymptotically 50% packing and covering optimal. Finally, greedy packing tends to choose many design points on the boundary of X. Relaxation can correspond to boundary avoidance, which in practice is shown to improve covering performance.

C0458: **Active sampling for high-dimensional ridge estimator with application in genome-wide association studies**
Presenter: **Lin Wang**, Purdue University, United States

Despite the availability of extensive data sets, it is often impractical to collect labels for all data points in many applications due to various measurement constraints. Subsampling approaches can be employed to select a subset of design points from a large pool, resulting in substantial savings in experimental costs. However, existing subsampling methods are primarily designed for low-dimensional data or rely on the assumption of sparse significant covariates. A computationally tractable sampling method is proposed that enables the selection of a small subset from a large data set without assuming sparsity. The method acknowledges the possibility that the number of significant covariates can be as large as or even larger than the sample size of the full data set. Specifically, the focus lies on ridge regression, for which sampling probabilities that minimize the mean squared predictive risk are developed on the full data set. The efficacy of the proposed approach is substantiated through theoretical analysis and extensive simulations. The results demonstrate its superiority over existing subsampling methods when dealing with high-dimensional data containing numerous significant covariates. Additionally, the advantages of the new approach are illustrated through its application to genome-wide association studies, highlighting its potential to yield valuable insights in this domain.

C0754: **A sequential approach to obtain optimal designs for non-linear models harnessing closed-form solutions**
Presenter: **Tirthankar Dasgupta**, Rutgers University, United States
Co-authors: Suvrojit Ghosh, Koulik Khamaru

D-Optimal designs for estimating parameters of response models are derived by maximizing the determinant of the Fisher information matrix, which depends on the unknown parameter vector of interest for non-linear models. Consequently, to obtain the D-optimal design for a non-linear model, one needs to have knowledge of the parameter to be estimated. One solution to this problem is to choose the design points sequentially, optimizing the D-optimality criterion using parameter estimates based on available data, followed by updating the parameter estimates using maximum likelihood estimation. On the other hand, there are many non-linear models for which closed-form results for D-optimal designs are available, but because such solutions involve the parameters to be estimated, they can only be used by substituting guestimates of parameters. A hybrid sequential strategy is proposed that replaces the optimization of the objective function at every single step by plugging in the estimates into the available closed-form solutions. Theoretical guarantees for the proposed approach are established. The usefulness of this approach in terms of saving computational time and achieving greater efficiency of estimation compared to the standard sequential approach is demonstrated with simulations conducted from two different sets of models motivated by real-life scenarios.

C0770: **Summary of effect aliasing structure for design selection and factor-column assignment for supersaturated designs**
Presenter: **Frederick Kin Hing Phoa**, Academia Sinica, Taiwan
Co-authors: Yi-Hua Liao, David Woods

In the assessment and selection of supersaturated designs, the aliasing structure of interaction effects is usually ignored in traditional criteria such as $E(s^2)$-optimality. A summary of the effect aliasing structure (SEAS) is introduced to assess the aliasing structure of supersaturated designs. SEAS takes into account the interaction terms and provides more informative summaries than traditional design criteria, such as (generalized) resolution and word length patterns, for design evaluations. The new summary consists of three criteria, abbreviated as MAP: (1) the maximum dependency aliasing (M-)pattern; (2) the average square aliasing (A-)pattern; and (3) the pairwise dependency ratio (P-)pattern. The relationships are theoretically studied among the three criteria of SEAS and traditional criteria, and the use of SEAS is demonstrated for evaluating and comparing some examples of supersaturated designs, including those suggested in the literature. The SEAS are further applied to the assignment of columns of a supersaturated design when some important experimental factors are known in prior.

CO355 Room K2.31 (Nash Lec. Theatre) STATISTICAL LEARNING IN CAUSAL INFERENCE Chair: Debarghya Mukherjee

C1042: **Learning the distribution map in reverse causal performative prediction**
Presenter: **Subha Maity**, University of Pennsylvania, United States

In numerous predictive scenarios, the predictive model affects the sampling distribution; for example, job applicants often meticulously craft their resumes to navigate through screening systems. Such shifts in distribution are particularly prevalent in the realm of social computing, yet the strategies to learn these shifts from data remain remarkably limited. Inspired by a microeconomic model that adeptly characterizes agents' behavior within labor markets, we introduce a novel approach to learning the distribution shift. The method is predicated on a reverse causal model, wherein the predictive model instigates a distribution shift exclusively through a finite set of agents' actions. Within this framework, a microfoundation model is employed for the agents' actions, and a statistically justified methodology is developed to learn the distribution shift map, which is demonstrated to be effective in minimizing the performative prediction risk.

C1083: **Controlling the false discovery proportion in observational studies with hidden bias**
Presenter: **Colin Fogarty**, University of Michigan, United States
Co-authors: Mengqi Lin

An approach to exploratory data analysis is proposed in matched observational studies. The setting where a single intervention is thought to potentially impact multiple outcome variables is considered, and the aim is to investigate which of these causal hypotheses come to bear while accounting not only for the possibility of false discoveries but also the possibility that the study is plagued by unmeasured confounding. For any candidate set of rejected hypotheses, the method provides sensitivity intervals for the false discovery proportion (FDP), the proportion of rejected hypotheses that are actually true. For a set containing L outcomes, the method describes how much unmeasured confounding would need to exist for us to believe that the proportion of true hypotheses is 0/L,1/L,..., all the way to L/L. Moreover, the resulting confidence statement intervals are valid simultaneously over all possible choices for the rejected set, allowing the researcher to look in an ad hoc manner for promising subsets of outcomes that maintain a large estimated fraction of correct discoveries even if a large degree of unmeasured confounding is present. The approach is particularly well suited to sensitivity analysis, as conclusions that some fraction of outcomes were affected by the treatment exhibit larger robustness to unmeasured confounding than the conclusion that any particular outcome was affected.

C1098: **Combining experimental and observational data for identification and estimation of long-term causal effects**
Presenter: **AmirEmad Ghassami**, Boston University, United States
Co-authors: Ilya Shpitser, Eric Tchetgen Tchetgen

The task of identifying and estimating the causal effect of a treatment variable is considered on a long-term outcome variable using data from an observational domain and an experimental domain. The observational domain is subject to unobserved confounding. Furthermore, subjects in the experiment are only followed for a short period of time; hence, long-term effects of treatment are unobserved, but short-term effects are observed. Therefore, data from neither domain alone suffices for causal inference about the causal and must be pooled in a principled way instead. Three approaches for data fusion are proposed. The first approach is based on assuming equi-confounding bias for the short-term and long-term outcomes. The second approach is based on a relaxed version of the equi-confounding bias assumption, where the existence of an observed confounder is assumed such that the short-term and long-term potential outcome variables have the same partial additive association with that confounder. The third approach is based on the existence of an extra variable in the system, which is a proxy of the latent confounder of the treatment-outcome relation. Influence function-based estimation strategies are proposed for each of the data fusion frameworks, and the robustness properties of the proposed estimators is studied.

C1316: **Efficiency and robustness of Rosenbaum's rank-based estimator in randomized experiments**
Presenter: **Nabarun Deb**, University of Chicago, United States
Co-authors: Bikram Karmakar, Aditya Ghosh, Bodhisattva Sen

Mean-based estimators of the causal effect in a completely randomized experiment may behave poorly if the potential outcomes have a heavy tail or contain outliers. An alternative estimator is studied by Rosenbaum that estimates the constant additive treatment effect by inverting a randomization test using ranks. By investigating the breakdown point and asymptotic relative efficiency of this rank-based estimator, it is shown that it is provably robust against outliers and heavy-tailed potential outcomes and has asymptotic variance at most 1.16 times that of the difference-in-means estimator (and much smaller when the potential outcomes are not light-tailed). A consistent estimator of the asymptotic standard error is further derived for Rosenbaum's estimator, which yields a readily computable confidence interval for the treatment effect. A regression-adjusted version of Rosenbaum's estimator is also studied to incorporate additional covariate information in randomization inference. The gain in efficiency is proven by this regression adjustment method under a linear regression model. It is illustrated through synthetic and real data that, unlike the mean-based estimators, these rank-based estimators (unadjusted or regression-adjusted) are efficient and robust against heavy-tailed distributions, contamination, and model misspecification. Finally, the study of Rosenbaum's estimator is initiated when the constant treatment effect assumption may be violated.

CO331 Room K2.41 COMPETING RISKS AND DEPENDENT SURVIVAL MODELS WITH PARAMETRIC ELEMENTS Chair: Dennis Dobler

C0407: **Estimation for the Mann-Whitney effect under parametric survival copula models**
Presenter: **Takeshi Emura**, The Institute of Statistical Mathematics, Japan

The Mann-Whitney effect is a measure for comparing survival distributions between two groups. The Mann-Whitney effect is interpreted as the probability that a randomly selected subject in a group survives longer than a randomly selected subject in the other group. Under the independence assumption of two groups, the Mann-Whitney effect can be expressed as the traditional integral formula of survival functions. However, when the survival times in the two groups are not independent of each other, the traditional formula of the Mann-Whitney effect has to be modified. A copula-based estimator is proposed for the Mann-Whitney effect with parametric survival models under the dependence of two groups, which may arise in the potential outcome framework. In addition, the standard error and confidence interval are derived based on the jackknife and asymptotic theory. Through a simulation study, the correctness of the proposed methods is shown. The proposed methods are applied to two real datasets.

C0528: **A comparison of Kaplan-Meier-based inverse probability of censoring weighted regression methods**
Presenter: **Morten Overgaard**, Aarhus University, Denmark

Weighting with the inverse probability of censoring is an approach to deal with censoring in regression analyses where the outcome may be missing due to right censoring. Three separate approaches involving this idea in a setting where the Kaplan-Meier estimator is used for estimating the censoring probability are compared. In more detail, the three approaches involve weighted regression, regression with a weighted outcome, and regression of a jack-knife pseudo-observation based on a weighted estimator. The asymptotic variance in each case is expressed, allowing for comparisons between each other and to the uncensored case. In terms of low asymptotic variance, a clear winner cannot be found. Which approach will have the lowest asymptotic variance depends on the censoring distribution.

C0613: **Testing similarity of parametric competing risks models for identifying potentially similar pathways in healthcare**
Presenter: **Kathrin Moellenhoff**, University of Cologne, Faculty of Medicine and University Hospital, Cologne, Germany, Germany
Co-authors: Nadine Binder, Holger Dette

The identification of similar patient pathways is a crucial task in healthcare analytics. A flexible tool to address this issue is parametric competing risks models, where transition intensities may be specified by a variety of parametric distributions, thus, in particular, being possibly time-dependent. The similarity between two such models is assessed by examining the transitions between different health states. A method is introduced to measure the maximum differences in transition intensities over time, leading to the development of a test procedure for assessing similarity. A parametric bootstrap approach is proposed for this purpose, and proof to confirm the validity of this procedure is provided. The performance of the proposed method is evaluated through a simulation study, considering a range of sample sizes, differing amounts of censoring, and various thresholds for similarity. Finally, the practical application of the approach is demonstrated with a case study from urological clinical routine practice, which inspired this research.

C1288: **Inference via wild bootstrap and multiple imputation under fine-gray models with incomplete data**
Presenter: **Marina Dietrich**, University Augsburg, MNTF, Prof. Friedrich, Germany
Co-authors: Dennis Dobler, Mathisca de Gunst

The wild bootstrap is a popular resampling method in the context of time-to-event data analysis. It can be used to justify the accuracy of inference procedures such as hypothesis tests or time-simultaneous confidence bands. In previous works, wild bootstrap confidence bands have been established for inference on cumulative incidence functions under the Fine-Gray proportional sub-hazards model if the data are censoring-complete. However, it is rather unusual that data are censoring-complete, and hence, this assumption is restrictive. In order to overcome this limitation, a novel wild bootstrap and multiple imputation-based (WB-MI) confidence band is proposed for the cumulative incidence function under the fine-gray model with incomplete data. Furthermore, the asymptotic validity of the proposed WB-MI confidence band is justified, and its reliability is numerically assessed and compared with already existing methods. The approach is illustrated by investigating the impact of pneumonia for intensive care unit patients on the probabilities of hospital death competing with live discharge.

CO022 Room S0.11 STATISTICS IN NEUROSCIENCE I Chair: Russell Shinohara

C0332: **Mapping individual differences in inter-modal coupling**
Presenter: **Sarah Weinstein**, Temple University, United States

Within-individual "coupling" between measures of brain structure and function evolves in development and may underlie differential risk for neuropsychiatric disorders. Despite increasing interest in the development of structure-function relationships, rigorous methods to quantify and test individual differences in coupling remain nascent. The purpose is to explore and address gaps in approaches for testing and spatially localizing individual differences in intermodal coupling. A new method, called CIDeR, is proposed and designed to simultaneously perform hypothesis testing in a way that limits false positive results and improves the detection of true positives. Through a comparison with existing approaches to quantifying and testing individual differences in coupling, subtle differences are delineated in their underlying null hypotheses, which may ultimately lead researchers to arrive at different results. Finally, the utility of CIDeR is illustrated in two applications to data from a large-scale study of neurodevelopment.

C0583: **A time-varying AR, bivariate DLM of functional near-infrared spectroscopy data**
Presenter: **Timothy Johnson**, University of Michigan, United States

Functional near-infrared spectroscopy (fNIRS) is a relatively new neuroimaging technique. It is a low-cost, portable, and non-invasive method to measure brain activity via the blood oxygen level-dependent signal. Similar to fMRI, it measures changes in the level of blood oxygen in the brain.

Its time resolution is much finer than fMRI. However, its spatial resolution is much closer, similar to EEG or MEG. fNIRS is finding widespread use on young children who cannot remain still in the MRI magnet, and it can be used in situations where fMRI is contraindicated, such as with patients who have cochlear implants. Furthermore, fNIRS measures the concentration of both oxygenated and deoxygenated hemoglobin, both of which may be of scientific interest. A fully Bayesian time-varying autoregressive model is proposed to analyze fNIRS data within the multivariate DLM framework. Low-frequency drift is modelled with a variable B-spline model (both locations and number of knots are allowed to vary). Both the model error and the auto-regressive processes vary with time. Simulation studies show that this model naturally handles motion artifacts and has good statistical properties. The model is then applied to a fNIRS data set.

C0677: **Multivariate inference for effect size images**
Presenter: **Simon Vandekar**, Vanderbilt University, United States

Inference in neuroimaging research has focused on hypothesis testing rather than estimation of effect sizes. As a solution, colleagues in biostatistics have developed procedures to construct spatial confidence sets for images that can be used to identify regions with target effect sizes above a given threshold with a specified probability. These confidence sets represent a paradigm shift in group-level inference for neuroimaging data. However, there is no generalized approach to estimating and constructing confidence regions on a unitless scale. This limits the general applicability of the approach. Nonparametric bootstrapping and recently developed approaches are used to construct confidence sets from simultaneous confidence intervals to establish a confidence set procedure for effect sizes of arbitrary model parameters. Their finite sample is evaluated, and the methods are used to identify regions associated with age and diagnostic differences in two datasets studying autism and psychosis.

C1150: **A case study of pupil dynamics after cannabis consumption using crossed multilevel function-on-scalar regression**
Presenter: **Julia Wrobel**, Emory University, United States

Marijuana is now legal for recreational or medical use in 41 states. Due to long-standing federal restrictions on cannabis research, the implications of cannabis legalization on traffic and occupational safety are understudied. Accordingly, there is a need for objective and validated measures of cannabis impairment that may be applied in public safety and occupational settings, such as post-crash or accident investigations. Identifying a reliable biomarker of recent cannabis use has proven challenging, but pupil response to light may offer an avenue for detection that outperforms typical sobriety tests. A wearable pupillometer is used to collect 5-second trajectories of change in pupil size after a light stimulus to investigate pupil light response as a biomarker of recent cannabis use. These "pupil trajectories" are collected for both the left and right eye before and at 45 and 80 minutes post-cannabis consumption for 120 subjects, resulting in 720 functional observations of pupil response to light. A new functional regression model is then developed to infer pupil dynamics in response to light for those who used and did not use cannabis and at different times post-cannabis use. The model modifies methodology on structured functional principal components analysis to provide appropriate inference for the complex repeated-measures structure of the functional observations. The method's performance is evaluated in simulations, and results for the motivating data are presented.

CO188 Room S0.12 TRANSFORMATIVE APPROACHES IN MACHINE LEARNING Chair: Chuan Hong

C1696: **Scalable Bayesian inference for the generalized linear mixed model**
Presenter: **Samuel Berchuck**, Duke University, United States
Co-authors: Sayan Mukherjee, Andrea Agazzi

The generalized linear mixed model (GLMM) is a popular approach for handling correlated data, and is used extensively in applications where big data is common, including biomedical settings. The focus is scalable statistical inference for the GLMM, where we define statistical inference as: (i) estimation of population parameters, and (ii) evaluation of scientific hypotheses in the presence of uncertainty. Artificial intelligence (AI) learning algorithms excel at scalable statistical estimation, but rarely include uncertainty quantification. In contrast, Bayesian inference provides full statistical inference, since uncertainty quantification results automatically from the posterior distribution. Unfortunately, Bayesian inference algorithms, including Markov Chain Monte Carlo (MCMC), become computationally intractable in big data settings. We introduce a statistical inference algorithm at the intersection of AI and Bayesian inference, that leverages the scalability of modern AI algorithms with guaranteed uncertainty quantification that accompanies Bayesian inference. Our algorithm is an extension of stochastic gradient MCMC with novel contributions that address the treatment of correlated data and proper posterior variance estimation. Through theoretical and empirical results, we establish our algorithm's statistical inference properties and apply the method to a large electronic health records database.

C1699: **Dynamic time-to-event prediction with ML/DL: Addressing competing risks in clinical outcomes**
Presenter: **Li Tang**, St. Jude Children's Research Hospital, United States

With the growing availability of longitudinal data from electronic medical records (EMR), integrating this data with baseline variables for dynamic prediction of clinically significant time-to-event outcomes has become a research priority. Previous studies have explored machine learning and deep learning models such as random survival forest, DeepSurv, and Transformers to incorporate multivariate longitudinal data into survival prediction. While these approaches show promise, the impact of competing risks common in survival analysis remains underexplored. Our study addresses this gap by evaluating these models under competing risks through simulations. We found that prediction accuracy is sensitive to landmark times, prediction windows, and interactions between variables. We then applied these refined methods to predict acute graft-versus-host disease (aGVHD) in pediatric oncology patients undergoing allogeneic hematopoietic cell transplantation. Using high-dimensional longitudinal EMR data, our approach significantly improves GVHD risk prediction, enabling early identification of high-risk patients and facilitating tailored prophylaxis strategies. We demonstrate the potential of AI-driven clinical risk prediction and underscore the need for careful application to ensure these tools truly enhance patient care.

C1700: **Interpretable machine learning-based scoring system for clinical decision making**
Presenter: **Nan Liu**, National University of Singapore, Singapore

There has been an increased use of scoring systems in clinical settings to assess risks in a convenient manner that provides important evidence for decision-making. Machine learning-based methods may be useful for identifying important predictors and building models; however, their 'black box' nature limits their interpretability as well as clinical acceptability. The aim is to introduce and demonstrate how interpretable machine learning can be used to create scoring systems for clinical decision-making.

C1698: **Multi-source stable variable importance measure via adversarial machine learning**
Presenter: **Molei Liu**, Columbia Mailman School of Public Health, United States

As part of enhancing the interpretability of machine learning, it is of renewed interest to quantify and infer the predictive importance of certain exposure covariates. Modern scientific studies often collect data from multiple sources with distributional heterogeneity. Thus, measuring and inferring stable associations across multiple environments is crucial in reliable and generalizable decision-making. We propose MIMAL, a novel statistical framework for Multi-source stable Importance Measure via Adversarial Learning. MIMAL measures the importance of some exposure variables by maximizing the worst-case predictive reward over the source mixture. Our framework allows various machine learning methods for confounding adjustment and exposure effect characterization. For inferential analysis, the asymptotic normality of our introduced statistic is established under a general machine learning framework that requires no stronger learning accuracy conditions than those for single source variable importance. Numerical studies with various types of data generation setups and machine learning implementation are conducted to justify the finite-sample performance of MIMAL. We also illustrate our method through a real-world study of Beijing air pollution in multiple locations.

CO264 Room S0.13 NEW ADVANCES IN BIOSTATISTICS Chair: Liqun Diao

C0685: **Analysis of gene expression data subject to measurement error in binary responses and predictors**
Presenter: **Li-Pang Chen**, National Chengchi University, Taiwan

Gene expression variables are usually used to classify specific diseases, such as acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL). However, gene expression data usually encounter ultrahigh dimensionality and measurement error. Those complex features also affect the classification performance. The aim is to introduce a method called BOOME, which refers to BOOsting algorithm for measurement error in binary responses and ultrahigh-dimensional predictors. This method primarily focuses on logistic regression and probit models with responses and predictors contaminated with measurement error. The BOOME method aims to address the effects of measurement error and then employs a boosting procedure to make variable selections and estimations. Numerical experiments reveal that the BOOME method is valid for addressing measurement error effects and deriving reliable estimation results.

C0727: **Zero-inflated Poisson models with measurement error in the response**
Presenter: **Qihuang Zhang**, McGill University, Canada

Zero-inflated count data arise frequently from genomics studies. Analysis of such data is often based on a mixture model, which facilitates excess zeros in combination with a Poisson distribution, and various inference methods have been proposed under such a model. Those analysis procedures, however, are challenged by the presence of measurement errors in count responses. A new measurement error model is proposed to describe error-contaminated count data. It is shown that ignoring the measurement error effects in the analysis may generally lead to invalid inference results, and meanwhile, situations are identified where ignoring measurement error can still yield consistent estimators. Furthermore, a Bayesian method is proposed to address the effects of measurement error under the zero-inflated Poisson model and discuss the identifiability issues. A data-augmentation algorithm is developed that is easy to implement. Simulation studies are conducted to evaluate the performance of the proposed method. The method is applied to analyze the data arising from a prostate adenocarcinoma genomic study.

C0830: **Diagnostic test accuracy meta-analysis based on exact within-study variance estimation method**
Presenter: **Zelalem Negeri**, University of Waterloo, Canada
Co-authors: Olana Dabi

Meta-analysis of diagnostic test accuracy studies commonly synthesizes study-specific test sensitivity and test specificity from different studies that aim to quantify the screening or diagnostic performance of a common index test of interest. A bivariate random effects model that utilizes the logit transformation of sensitivity and specificity and accounts for the within- and between-study heterogeneity is commonly used to make statistical inferences about the unknown test characteristics. However, it is well reported that this model may lead to misleading inference since it employs the logit transformation and approximate within-study variance estimate. Alternative transformations which do not require continuity corrections, such as the arcsine square root and Freeman-Tukey double arcsine, were recently proposed to overcome the former limitation. However, these solutions also suffer from using approximate within-study variance estimates, which can only be justified when within-study sample sizes are large. To overcome these problems, an exact within-study variance estimation approach is proposed, which does not require a continuity correction and is invariant to transformations. The new method and the existing approaches are evaluated using real-life and simulated meta-analytic data.

C1136: **Integrating probability and non-probability samples with misclassified covariates**
Presenter: **Hua Shen**, University of Calgary, Canada

In statistics, probability samples often lack crucial, precisely measured outcome variables, whereas non-probability samples provide detailed data but are not representative. Both sample types can also involve covariates subject to misclassification. A novel two-stage estimation technique is introduced to integrate these sample types, addressing misclassification in categorical covariates using a latent-variable framework without requiring validation data. Simulations demonstrate that the method achieves superior accuracy compared to traditional approaches. Furthermore, its effectiveness is validated on a real dataset, enhancing inference quality in studies with misclassified covariates.

CO348 Room Safra Lec. Theatre FUNCTIONAL DATA ANALYSIS AND STOCHASTIC PROCESSES Chair: Caterina May

C0513: **A nonparametric method to detect the number of components for dimensionality reduction techniques**
Presenter: **Enea Bongiorno**, Universita del Piemonte Orientale, Italy
Co-authors: Kwo Lik Lax Chan, Aldo Goia, Philippe Vieu

Working with data in high-dimensional or infinite-dimensional spaces often necessitates the use of dimensionality reduction techniques. One of the primary challenges is determining the number of components (or dimensions) to consider. A nonparametric technique is proposed to address this issue based on the concepts of complexity and small-ball probability. The method is demonstrated through examples and practical applications.

C0517: **Exponential expansions for approximation of probability distributions**
Presenter: **Anna Maria Gambaro**, Universita del Piemonte Orientale, Italy

Polynomial-based expansions have been widely used in literature to approximate probability density functions. A notable example is the A-type Gram-Charlier (AGC) expansions. The AGC expansion does not guarantee the positiveness of the truncated series, which, therefore, does not constitute a valid probability density function (PDF). To encompass the above drawback, prior studies propose to approximate PDF using the C-type Gram-Charlier (CGC) expansion. This expansion is of an exponential form that always guarantees positive truncated probabilities for any degree of skewness and kurtosis of the true PDF. The existing literature is extended in two main directions. Firstly, a promising link is highlighted between the exponential expansion approach for approximating PDFs and the theory of Bayes spaces, which is extensively studied in functional statistics. In particular, novel findings are introduced concerning the convergence of this series towards the true density function, employing mathematical tools of Bayes spaces. Secondly, the moment-based estimation of the coefficients of the exponential expansion is studied. A simple linear system is proposed to estimate the expansion coefficients, given the first n exact moments of the corresponding distributions and for any orthogonal polynomial basis of the Bayes space. Finally, numerical examples are provided that effectively demonstrate the efficiency and straightforward implementability of the proposed approach.

C0866: **Anomaly detection in profile monitoring using functional conformal prediction**
Presenter: **Simone Vantini**, Politecnico di Milano, Italy
Co-authors: Teresa Bortolotti, Egon Prioglio, Bianca Maria Colosimo

A novel methodology is proposed for detecting anomalies in the monitoring of profiles in industrial processes. The approach is grounded in functional data analysis (i.e., by modeling profiles as functions) and integrates conformal prediction and copula-based methods to detect unusual patterns. Theoretically, the control of the probability of having one or more false anomalies (i.e. type I errors) along the functional domain is also guaranteed in the case of small sample sizes and non-Gaussian data, which are common, for instance, in 3D printing applications. The functional control limits are obtained by inverting simultaneous functional conformal prediction bands, which have been proposed in the literature recently. Furthermore, to enhance interpretability and increase the anomaly detection power of the proposed procedure, the methodology is extended by employing copulas to simultaneously monitor functions and their higher-order derivatives. An extensive simulation study showcases the potential of the proposed approach and proves the effectiveness of functional conformal prediction and copula adjustment in detecting anomalies while controlling the probability of false anomalies. The applicability of the methodology is finally illustrated across various industrial applications in the field of 3D printing.

C1141: **Block testing covariance and precision matrices for functional data analysis**
Presenter: **Alessia Pini**, Universita Cattolica del Sacro Cuore, Italy
Co-authors: Marie Morvan, Joyce Madison Giacofci, Valerie Monber

A method to test linear independence and conditional linear independence is proposed between portions of the domain of functional data. Data is assumed to be described by means of a B-splines basis expansion, such that coefficients of the basis expansion are directly related to the parts of the domain where the support of basis functions is strictly positive. The domain is further assumed to be partitioned into regions of interest. In such a case, the precision matrix is expected to have a block structure, where blocks correspond to elements of the partition. To infer which areas of the domain are conditionally independent of each other, a permutation test is proposed on blocks of the covariance or precision matrix of basis coefficients. A suitable strategy is introduced to deal with the multiple testing issues in this setting. It is shown that the procedure can identify the true structure of dependence on simulated data and on a real case study involving tractographic data related to the infrared emission spectra of fruit purees.

CO298 **Room BH (S) 1.01 Lec. Theatre 1** **FORECASTING AND PANEL ANALYSIS** **Chair: Peter Pedroni**

C0162: **Forecasting after the start of a trend break**
Presenter: **David Hendry**, University of Oxford, United Kingdom
Co-authors: Jennifer L Castle, Jurgen Doornik

A sequence of large same-sign 1-step-ahead forecast errors occurring as the forecast origin advances signal a sudden location or trend shift. To test which, impulse indicators are included as intercept corrections to offset each forecast error, then replaced after 2 or 3 significant outcomes by a broken linear or log-linear trend, tested against each other and a step indicator. The approach is illustrated by a rapid increase in global mean sea level after 2011 quickly detected, providing accurate forecasts a decade ahead and accurately forecasting the UK's inflation over 2021–24 as compared with Bank of England projections. Simulations and analyses confirm the general feasibility of detecting sudden trend changes after just 2 or 3 periods and then forecasting reasonably accurately.

C1697: **Firm-level markups and monetary policy transmission: A panel time-series based analysis**
Presenter: **Utsav Bahl**, Williams College, Switzerland

Theoretical and empirical economics suggest that market power affects macroeconomic dynamics; firms with substantial market power exhibit less sensitivity to downturns and policy shifts. Moreover, these firms can manipulate prices, influence wages, and control market supply. This should, in theory, insulate them from immediate economic pressure, including monetary policy shocks. Additionally, sustained periods of accommodative monetary policy might enhance market power by enabling larger firms to leverage lower borrowing costs to consolidate their market position. The dynamic relationship between firms with high markup abilities and their response to monetary policy shocks is analyzed. In doing so, the policy implications of results are considered. Using Compustat data from over 100,000 firms across the US and globally, Heterogeneous Panel Structural Vector Autoregression and Heterogeneous Functional Structural Vector Autoregression approaches are used to allow for a granular examination of how firms with varying levels of market power respond to monetary shocks and how these responses aggregate to influence broader macroeconomic outcomes.

C1663: **Automatic stabilizers from fiscal policy and the role of public and private consumption**
Presenter: **Peter Pedroni**, Williams College, United States
Co-authors: Fakhri Hasanov

A large panel of quarterly data observed over 52 countries from 1980 to 2023 is used to study the stabilization role of different forms of fiscal policy that operate via the dynamic interaction between public and private consumption. In particular, a heterogeneous panel VAR approach is employed, which allows the usage of standard structural identifying assumptions from the macro literature to disentangle exogenous discretionary fiscal policy shocks from the endogenous automatic budgetary responses of fiscal policy to economic shocks. In this context, measures are developed that allow the investigation via counterfactual historical decompositions of the extent to which exogenous versus endogenous fiscal policy has served to stabilize macroeconomic fluctuations in GDP and private consumption. Cross-country heterogeneity is also exploited in these measures to explore which economic structural attributes have been associated with greater or lesser stability outcomes for exogenous versus endogenous fiscal policy. It is found that in addition to the usual factors such as exchange rate regimes and degrees of openness, the presence of sovereign wealth funds and debt-based fiscal rules also play an important role.

C1660: **Innovation in solar energy technologies and the implications for environmental and economic sustainability**
Presenter: **Fakhri Hasanov**, KAPSARC, Saudi Arabia
Co-authors: Peter Pedroni

The focus is on investigating the extent to which various technological innovations in solar energy have facilitated climate change mitigation and economic development as measured by CO2 emissions and GDP via the generation of renewable energy. It is done in the context of a heterogeneous panel VAR analysis for which a multi-country panel of annual data on solar energy patents spanning from 1990 to 2021 is employed. The approach also allows the exploration of the extent to which various policies have tended to enhance the link between innovations in solar energy technologies and sustainable environmental and economic outcomes. Examples explored include the degree of international integration with respect to global value chains, foreign direct investment and general trade openness. Other examples include spending on human capital development and environmental regulatory policies. It also demonstrates how the approach can be used to study the experiences of individual countries and include case studies for prominent energy-producing countries such as Saudi Arabia. Finally, it is shown how to incorporate the results of the analysis into large-scale macro-econometric models such as Saudi Arabia's KGEMM energy and emissions augmented macro-econometric model with an eye toward informing the design of policies to achieve and maintain sustainability in energy, economic development and the environment in line with overall development goals.

CO398 **Room BH (SE) 1.01** **RECENT ADVANCES IN BAYESIAN METHODS** **Chair: Luca Aiello**

C0319: **Model based clustering of time-dependent observations with common historical shocks**
Presenter: **Luca Danese**, University of Milano-Bicocca, Italy
Co-authors: Andrea Ongaro, Riccardo Corradin, Wasiur Rahman KhudaBukhsh

A novel model-based clustering approach is proposed for samples of time series. A unique commonality is assumed that two observations belong to the same group if structural changes in their behaviors happen at the same time. A latent representation of structural changes is resorted to in each time series based on random orders to induce ties among different observations. Such a general approach can be combined with many time-dependent models known in the literature. The motivation is the epidemiological problem, where the aim is to provide clusters of different countries of the European Union, where two countries belong to the same cluster if the diffusion processes of the COVID-19 virus had structural changes at the same time.

C0579: **Dynamic mixture of finite mixtures of factor analyzers**
Presenter: **Margarita Grushanina**, Imperial College London, United Kingdom
Co-authors: Sylvia Fruehwirth-Schnatter

Mixtures of factor analyzers represent a popular tool for finding structure in data. While in many applications, the number of clusters and latent factors within clusters is held constant, some recent models automatically infer cluster and/or factor dimensionalities. This is done by employing nonparametric priors and allowing the number of clusters and factors to potentially be infinite. MCMC estimation is performed via adaptive algorithms, where parameters associated with the redundant factors are discarded. The current work contributes to the literature by allowing automatic inference on the number of clusters and cluster-specific factors while keeping both dimensions finite. For automatic inference on the cluster structure, the dynamic mixture of finite mixtures model is employed with a prior on the number of mixture components. Automatic inference on cluster-specific factors is performed by assigning an exchangeable shrinkage process (ESP) prior, which can be interpreted as a generalized cumulative shrinkage process (CUSP) prior to the columns of the factor loading matrices. Extensive simulation studies and applications to benchmark as well as real data sets demonstrate that the model outperforms competing alternatives, in particular, based on the multiplicative gamma process prior, with respect to recovering the correct number of cluster-specific factors.

C0776: **"Slice" the tables in the hierarchical Dirichlet process**
Presenter: **Matteo Gianella**, Politecnico di Milano, Italy
Co-authors: Mario Beraha, Federico Camerlenghi, Alessandra Guglielmi

The hierarchical Dirichlet process (HDP) is a well-known and well-established nonparametric Bayesian model that extends the Dirichlet process to handle partially exchangeable data. It is particularly useful in scenarios where data are naturally organised into groups, and the statistical target is to identify shared clusters across these groups. Since its introduction back in 2004, the HDP has been consistently used as a mixing measure in parametric mixtures for clustering in various contexts, e.g., in topic modelling, collaborative filtering and gene expression analysis. Posterior inference in the HDP is typically based on the so-called Chinese restaurant franchise metaphor, whereby customers sit at different tables as in the original Chinese restaurant process, but different tables might share the same dish. This makes posterior inference cumbersome as one must introduce two sets of latent indicator variables that interact. The aim is to demonstrate how, by simply assuming a prior distribution on the concentration parameter of the HDP, another representation of the HDP can be obtained. In this new representation, all tables serve distinct dishes, paving the way for a slice sampling algorithm that reduces the amount of bookkeeping required and significantly improves posterior computation in terms of efficiency.

C0851: **Multivariate treatment effect estimation through Bayesian factor regression model**
Presenter: **Dafne Zorzetto**, Brown University, United States

In the context of causal inference, investigating causal effects for multivariate potential outcomes has not yet been extensively explored. This is due to the missing data problem inherent in the potential outcomes framework, which leads to the challenges of capturing the causal effect of a treatment on correlated outcomes and understanding the heterogeneity of the causal effect between outcomes. The ability of Bayesian factor analysis is exploited to identify the latent treatment-specific factors that capture and characterize the causal effects within correlated multivariate outcomes. The innovative use of the dependent Dirichlet process as the distribution for the factor scores allows the overcoming of the problem of missing data through an aware and fair imputation. Aligned to real data questions in environmental epidemiology, the causal link between air pollution regulation and the concentration of various pollutants in the United States is investigated.

CO098 Room BH (SE) 1.02 BAYESIAN SEMI- AND NON-PARAMETRIC METHODS III Chair: Guillaume Kon Kam King

C0819: **Group sparse variational Bayes approach for high-dimensional data**
Presenter: **Sarah Filippi**, Imperial College London, United Kingdom
Co-authors: Michael Komodromos, Kolyan Ray, Marina Evangelou

Few Bayesian methods for analyzing high-dimensional sparse data provide scalable variable selection, effect estimation and uncertainty quantification. Most methods either sacrifice uncertainty quantification by computing maximum a posteriori estimates or quantify the uncertainty at high (unscalable) computational expense. The focus is on a method for the selection of groups of variables under a generalized linear model. For this setting, an interpretable and scalable model is developed for prediction and group selection. The method, based on a variational approximation, overcomes the high computational cost of MCMC whilst retaining the useful features, providing excellent point estimates and offering a natural mechanism for group selection via posterior inclusion probabilities. Posterior concentration rates are derived for the group sparse variational Bayes approach. They compare the methods against other state-of-the-art Bayesian variable selection methods on simulated data and demonstrate their application for variable and group selection on real biomedical data.

C0318: **Density regression via Dirichlet process mixtures of normal structured additive regression models**
Presenter: **Vanda Inacio**, University of Edinburgh, United Kingdom

Within Bayesian nonparametrics, dependent Dirichlet process mixture models provide a flexible approach for conducting inference about the conditional density function. However, several formulations of this class make either restrictive modelling assumptions or involve intricate algorithms for posterior inference. A flexible and computationally tractable approach is presented for density regression based on a single-weights dependent Dirichlet process mixture of normal distributions model for univariate continuous responses. An additive structure is assumed for the mean of each mixture component, and the effects of continuous covariates are incorporated through smooth functions. The key components of the modelling approach are penalised B-splines and their bivariate tensor product extension. The method also seamlessly accommodates categorical covariates, linear effects of continuous covariates, varying coefficient terms, and random effects, which is why the model is referred to as a Dirichlet process mixture of normal structured additive regression models. Results from a simulation study demonstrate that the approach successfully recovers the true conditional densities and other regression functionals in challenging scenarios. Applications to three real datasets further underpin the broad applicability of the method.

C0331: **Bayesian mixture models for histograms: With applications to large datasets**
Presenter: **Richard Warr**, Brigham Young University, United States

It is not uncommon for privacy or summarization purposes to receive data in a table or in histogram format with bins and associated frequencies. A method that fits a mixture distribution to model the probability density function of the underlying population is presented. The focus is on a mixture of normal distributions, however the method could be generalized to mixtures of other distributions. A prior is placed on the number of mixture components, which could be finite or countably infinite and inference is obtained using reversible jump MCMC. The attractive properties of the method are demonstrated, and they show a great deal of promise for modeling large data problems using a Bayesian nonparametric approach. Additionally, multiple histograms are considered, and those are clustered using the Dirichlet process. This clustering allows for the sharing of information between populations and provides a posterior probability of homogeneity between populations.

C1375: **Graph of graphs: From nodes to supernodes in graphical models**
Presenter: **Alexandros Beskos**, University College London, United Kingdom
Co-authors: Maria De Iorio, Willem van den Boom, Ajay Jasra, Andrea Cremaschi

High-dimensional data analysis typically focuses on low-dimensional structure, often to aid interpretation and computational efficiency. Graphical models provide a powerful methodology for learning the conditional independence structure in multivariate data by representing variables as nodes and dependencies as edges. Inference is often focused on individual edges in the latent graph. Nonetheless, there is increasing interest in determining more complex structures, such as communities of nodes, for multiple reasons, including more effective information retrieval and better interpretability. A hierarchical graphical model is proposed where nodes are first clustered, and then, at a higher level, the relationships are investigated among groups of nodes. Specifically, nodes are partitioned into supernodes with a data-coherent size-biased tessellation prior, which combines ideas from Bayesian nonparametrics and Voronoi tessellations. This construct also allows accounting for the dependence of nodes within supernodes. At the higher level, the dependence structure among supernodes is modeled through a Gaussian graphical model, where the focus of inference is on superedges. Theoretical justification is provided for the modeling choices. Tailored MCMC schemes are designed to enable parallel computations. The effectiveness of the approach is demonstrated for large-scale structure learning in simulations and a transcriptomics application.

CO387 Room BH (SE) 1.05 NOWCASTING AND FORECASTING MACROECONOMIC AND FINANCIAL RISK Chair: Domenico Giannone

C0237: Measuring the Euro area output gap

Presenter: **Matteo Luciani**, Federal Reserve Board, United States

Co-authors: Matteo Barigozzi, Claudio Lissona

The Euro area output gap and potential output are measured using a non-stationary dynamic factor model estimated on a large dataset of macroeconomic and financial variables. It is found that after the 2008 global financial crisis, potential output growth slowed, and, as of the end of 2023, it has yet to return to its pre-2008 pace. As a result, it is estimated that from 2012 to 2023, the economy was tighter than estimated by the European Commission and the IMF. Moreover, credit indicators are crucial for capturing the business cycle's medium-term fluctuations, thus confirming that the business and financial cycles are correlated and co-move in the medium run.

C0718: Understanding growth-at-risk: A Markov-switching approach

Presenter: **Francesca Loria**, Federal Reserve Board, United States

Both financial and macroeconomic conditions matter for downside risks to the economic outlook. It is shown that the deterioration of the financial and real sides dramatically increases the probability of tail risks of large negative growth over the next year. A real-time measure of financial conditions and economic activity is proposed for the United States, and these measures are used to construct conditional quantiles and predictive distributions of average GDP growth over the next 12 months. It is found that periods of high macro and financial distress, such as the global financial crisis and the COVID-19 pandemic, are associated with low average future growth, high uncertainty, and risks tilted to the downside. This methodology is a powerful tool to assess the risk of tail events, such as recessions, and to evaluate the likelihood of point forecasts.

C0956: Forecasting macroeconomic risk with many predictors

Presenter: **Domenico Giannone**, University of Washington, United States

A quantile regression is developed for big data using shrinkage and deep neural networks to forecast macroeconomic risk using a wide array of predictors.

C1201: Market returns dormant in options panels

Presenter: **Yoosoon Chang**, Indiana University, United States

Co-authors: Soohun Kim, Joon Park, Youngmin Choi

The purpose is to uncover the relationship between extensive option panels and market returns through functional predictive regression. Employing the approach on the realized returns of the S&P 500, remarkable performance is achieved in predicting S&P 500 monthly returns, yielding an in-sample R-squared of 5.12% and an out-of-sample R-squared of 4.88%. Additionally, the method proves highly effective in predicting the realized variance of the S&P 500 index, achieving in-sample and out-of-sample R-squared values of 33.65% and 23.35%. The predictive accuracy of the model surpasses that of established predictors and equilibrium modes. It is found that utilizing the risk-neutral density as a predictor and employing the functional regression approach is indispensable for achieving this level of outperformance.

CO049 Room BH (SE) 1.06 CLIMATE AND SUSTAINABLE FINANCE Chair: Monica Billio

C1179: The incremental information content of ESG score in financial decision making

Presenter: **Michele Costola**, Ca' Foscari University of Venice, Italy

The purpose is to examine the incremental information content of ESG scores for investors and decision-making beyond traditional financial metrics such as revenue, profitability, debt, and financial performance indices. Whether ESG variables capture information not reflected in market valuations or demographic characteristics is investigated, including firm size, industry, and geographic location. The analysis aims to shed light on the relevance of ESG ratings for regulators, portfolio managers, and firms, with potential implications for their materiality.

C1533: ESG-constrained portfolio choice with estimation risk

Presenter: **Chengguo Weng**, University of Waterloo, Canada

Environmental, social, and governance (ESG) investing has emerged as a global trend, offering sustainable benefits to investors and financial institutions. An ESG constraint is integrated into the classical mean-variance optimization framework while accounting for the estimation risk associated with the first two moments of asset returns. It begins by examining the problem in the absence of estimation risk and deriving the optimal portfolio characterized by three-fund separation. To address estimation risk, a combined three-fund portfolio is proposed, with components based on a plug-in ESG portfolio. The optimal combination coefficients are derived by maximizing the expected out-of-sample mean-variance utility. Extensive simulations and empirical analysis demonstrate that the combined portfolio outperforms the plug-in ESG portfolio in terms of certainty-equivalent return. Furthermore, a comparative performance analysis is provided between the two ESG-constraint types involved.

C1593: Hydrogeological risk and accessing credit for Italian SME

Presenter: **Fabio Parla**, University of Palermo, Italy

Co-authors: Andrea Cipollini, Fabio Parla

The focus is on the impact of hydrogeological risk on credit market conditions for Italian SME. The contribution to the recent literature, which uses NUTS3 data for Europe, is by using a more granular dataset computed at the municipal level. More specifically, the reliance is on a Panel VARX fitted to proxies of credit market conditions retrieved from European DataWarehouse (EDW). The exogenous variable is the climate sentiment time series for Italy, interacting with cross-sectional data measuring hydrogeological risk available at the municipal level from ISPRA. The two endogenous variables are, first, two different proxies of access to finance: either interest rates on loans or a credit supply indicator constructed using the procedure developed by a past study. The other endogenous variable is the default rate. The quarterly dataset available for the 2008-2021 period allows the usage of a mean group estimator to account for heterogeneity. As a robustness check, the size of the bank granting loans and the firm accessing credit is controlled for. Moreover, the loss given default is computed at the municipal level as an alternative proxy of credit risk. While the literature using proxies of credit market conditions as dependent variables in single equation regression implicitly considers the endogenous variables orthogonal to each other, the use of a Panel VAR allows accounting for their interaction, avoiding a downward bias in stress testing exercises.

C0444: Mutual funds' appetite for sustainability in European auto ABS

Presenter: **Carmelo Latino**, Leibniz Institute for Financial Research SAFE, Germany

Co-authors: Yue Wang, Loriana Pelizzon, Max Riedel

The purpose is to study how mutual funds contribute to financing the transition to zero- or low-emission vehicles (ZLEVs) in Europe. Hand-collected data uncover three novel types of sustainability measures for auto asset-backed securities (auto ABS) based on prospectus-, loan-, and manufacturer-level sustainability information. It is found that green funds tend to tilt their exposure to sustainability-transparent auto ABS and invest marginally more in deals with a higher proportion of ZLEVs in the underlying collateral pool. However, in the absence of a globally accepted framework for green securitizations, asset managers use sustainability proxies that are associated with the lowest disclosure processing costs.

CO184 Room BH (S) 2.01 ADVANCES IN TIME SERIES AND PANEL DATA ECONOMETRICS Chair: Martin Wagner

C1241: Long-run money demand reconsidered

Presenter: **Sebastian Veldhuis**, University of Klagenfurt, Austria

Co-authors: Martin Wagner

Motivated by widely reported instabilities of money demand functions, a systematic analysis of potential sources and forms of such instabilities is performed. By means of a systematic econometric analysis that includes moving window analysis, formal monitoring procedures as well as nonlinearity testing of linear (cointegrating) relationships, the attempt is to obtain an improved understanding concerning instabilities, in particular, whether the sources of instabilities are comparable across countries and/or over different periods (monetary regimes). With respect to modeling nonlinear demand functions, the focus is on smooth transition cointegration analysis, with an implied focus on observable drivers of instabilities - or, more generally, of nonlinearities. The analysis is performed on a newly constructed annual frequency dataset that includes (depending upon country) real GDP, a broad and narrow monetary aggregate, consumer price and GDP deflators, a short- and long-term interest rate, the unemployment rate as well as a stock market index for about 20 countries for (depending upon country as early as) 1870 to 2023.

C1183: Online breakpoint - detection in cointegrating relationships

Presenter: **Leopold Soegner**, Institute for Advanced Studies, Austria

Co-authors: Martin Wagner

A closed-end consistent monitoring procedure is developed with the goal of detecting structural changes in cointegrating relationships. A vector error correction model is considered and allows for different specifications of the deterministic terms. A monitoring test statistic is proposed to investigate the stability of cointegrating relationships. The asymptotic distribution of the test statistic is obtained under the null hypothesis of no structural breaks. A calibration period is used for parameter estimation, after which online break-point detection is performed. The procedure stops at the first time point when the test statistic exceeds the corresponding critical value. A simulation study is provided to investigate the finite sample properties of our monitoring procedure.

C1167: FM-OLS estimation and inference for SUCPRs with common integrated regressors

Presenter: **Martin Wagner**, University of Klagenfurt, Bank of Slovenia and Institute for Advanced Studies, Vienna, Austria

Co-authors: Fabian Knorre

Two fully modified OLS (FM-OLS) type estimators are developed for systems of seemingly unrelated cointegrating polynomial regressions with common regressors, i.e., systems of regressions that include deterministic variables, integrated processes, integer powers of integrated processes as well as common - across (potentially subsets of) equations - integrated processes and integer powers of common integrated processes as explanatory variables. The stationary errors are allowed to be serially correlated, and the regressors to be endogenous. Furthermore, the errors and regressors are allowed to be dynamically cross-sectionally correlated. The developed estimators have zero mean Gaussian mixture limiting distributions that allow for asymptotic normal or chi-squared inference. The Wald-type hypothesis tests are used as a basis to formulate tests for general forms of group-wise poolability. In case group-wise poolability is not rejected, the corresponding group-wise pooled variants of the developed FM-OLS-type estimators are provided. The simulations indicate that appropriate pooling leads, as expected, to improved performance of both the estimators and hypothesis tests based upon them. The developed methodology is applied to analyzing the environmental and material Kuznets curve hypotheses for multi-country data for CO2 and SO2 emissions and aluminum, lead and zinc usage.

C0234: Noise cancelling observation-driven models

Presenter: **Jannik Steenbergen**, Aarhus University, Denmark

Co-authors: Leopoldo Catania

A new class of observation-driven models are proposed with a filter that downweighs the forcing variable when the model error term is below a certain threshold in absolute magnitude. The new updating mechanism relies on the premise that error terms of low absolute magnitude generally constitute noise and do not indicate a change in the time-varying parameter. The asymptotic properties of the maximum likelihood estimator of static model parameters are established, and the usefulness of the new noise-cancelling updating mechanism is demonstrated in a US inflation forecasting study. Results suggest that the new updating mechanism can improve the accuracy of forecasts compared to standard observation-driven counterparts.

CO390 Room BH (S) 2.02 NEW DIRECTIONS IN DETECTING CHANGES AND CLUSTERS Chair: Ansgar Steland

C1235: Asymptotic properties of k-means and its bias correction under high dimensional settings

Presenter: **Kento Egashira**, Tokyo University of Science, Japan

Co-authors: Kazuyoshi Yata, Makoto Aoshima

K-means is widely regarded as an effective method for analyzing high-dimension, low-sample-size (HDLSS) data, yet its asymptotic properties in such settings remain underexplored. The aim is to delve into the behavior of k-means in practical settings, even in the context of high-dimensional data. Based on these findings, a bias-corrected version of k-means is proposed to enhance performance. Additionally, the lack of theoretical analysis is addressed for kernel k-means in high-dimensional settings by examining its properties using a Gaussian kernel. This allows for investigating the theoretical comparison between kernel k-means and standard k-means. Numerical simulations demonstrate the effectiveness of the proposed bias-corrected k-means and conventional k-means, including kernel k-means, in high-dimensional contexts.

C1368: Optimal kernel smoothing method for detecting fixed and random mean change in multivariate data

Presenter: **Yanhong Wu**, California State University Stanislaus, United States

To bolster the effectiveness of sequential detection methods, a kernel smoothing moving average (KSMA) chart is introduced that emphasizes recent observations. This novel approach is pitted against established methods like the finite moving average (FMA), the exponentially weighted moving average (EWMA), and the CUSUM charts, scrutinized through the lens of the conditional average delay detection time (ADDT) for a given average in-control run length (ARL0) in scenarios of persistent change, as well as the percentiles for transient change. A tailored kernel function aimed at optimizing efficiency within the spectrum of signal strength is proposed. The comparative analysis spans both fixed mean and random mean change models, encompassing both univariate and multivariate observations. Findings underscore that the proposed method not only outperforms the CUSUM and Shiryayev-Roberts (S-R) procedures at the specified signal strength reference value but also surpasses the performance of EWMA and FMA charts. Detection of both the fixed and random mean change in daily returns for Dow Jones Industrial Stocks is used for illustration.

C1572: General adapted-threshold monitoring in discrete environments and rules for imbalanced classes

Presenter: **Ansgar Steland**, RWTH Aachen University, Germany

A framework for monitoring is studied where observed variables X_t are thresholded to detect relevant changes. In the novel approach, the threshold depends on external information Z_t, which is used to control the detector's sensitivity. As an example, consider individualized intense-care monitoring, where clinically relevant alarm thresholds need to depend on a patients' health status. Discrete-valued Z_t is studied, which splits the sample into classes, and threshold functions of Z_t are derived, controlling the false alarm rate α. A proportional threshold is proposed, which favors classes with small class probabilities, thus mitigating the imbalanced classes problem. Asymptotic theory is developed for i.i.d. and dependent learning samples. Two-stage designs are suggested, allowing the distribution of the budget in a controlled manner over an a priori partition of the sample space of Z_t.

C1356: **Statistical inferences from biomechanical fatigue data of running athletes**
Presenter: **Rupsa Basu**, University of Cologne, Germany

Statistical methodologies relevant to the study of biomechanical fatigue data from runners are explored. In particular, the focus is on the hip, knee and ankle angle data collected from both novice as well as experienced runners during a fatiguing protocol. In the biomechanical world, it is understood that fatigue induces specific changes in the movement of these joints during the course of the run. These changes are a consequence of the body adjusting to tiring conditions. Some of these adaptations of the body may, however, result in long-term injuries. It is, therefore, essential to detect these changes and make appropriate adjustments to the training strategies. In this regard, various statistical methodologies developed specifically for this data example are assessed. From the statistical aspect, relevant size change point detection for functional data is studied. The methodologies developed therein contribute to retrospective change analysis of biomechanical data. Further, online change analysis is also studied using a martingale statistic to enable biomechanical scientists to study real-time fatigue while the athlete is still running. Overall, it is shown how the methods presented provide biomechanical scientists and sports trainers with methodologies to understand the individual movement patterns of running athletes.

CO326 Room BH (S) 2.03 NEW DEVELOPMENTS IN FINANCIAL TIME SERIES Chair: Richard Gerlach

C0412: **Interval-based time series analysis: Detecting structural shifts and estimating change-points**
Presenter: **Li-Hsien Sun**, National Central University, Taiwan

A novel method is presented for detecting structural shifts within interval-based time series data and accurately estimating the change point. The approach involves developing a financial time series model that considers daily maximum, minimum, and terminal values based on the geometric Brownian motion model. The proposed method is then applied to the financial time series where the data is emphasized by the significance of intra-daily information, including maximum and minimum prices, while traditional finance models primarily focus on the daily closing price given the open price. The likelihood function and corresponding maximum likelihood estimates (MLEs) are derived using the Girsanov theorem and the Newton-Raphson (NR) algorithm. Through extensive simulations, the effectiveness of the proposed approach is thoroughly evaluated. Furthermore, empirical studies using real stock return data (S&P 500 index) from two critical periods, the 2008 financial crisis, the COVID-19 pandemic in 2020, and the Russo-Ukrainian War in 2022, allow assessing its performance robustly.

C0415: **Nonlinear market dynamics: Range-based hysteretic volatility modeling**
Presenter: **Edward Meng-Hua Lin**, Tunghai University, Taiwan

The hysteretic threshold conditional autoregressive range model (HTARR model) is introduced, integrating the hysteretic structure into the range-based volatility models. By using the proposed model, nonlinear structures that are more consistent with changes in market volatility are captured, and the performance necessary for effective market volatility prediction is obtained. Bayesian methods are employed to estimate the model's unknown parameters, and simulation studies are conducted to validate its feasibility. Furthermore, the model's performance is evaluated and compared with other volatility models through the analysis and prediction of intra-day range data from eight financial markets.

C0435: **Portfolio optimization based on dynamic networks and vine copulas**
Presenter: **Shih-Feng Huang**, National Central University, Taiwan

The application of vine copulas combined with network methods is explored for portfolio optimization. It begins by eliminating inherent features such as autocorrelation, conditional heteroscedasticity, and volatility clustering in each financial time series using the de-GARCH technique. The similarity matrix of the multivariate de-GARCH series is then calculated to construct the global minimum spanning tree (MST), which helps identify suitable stocks for the portfolio. Subsequently, the local MST (LMST) is built for the selected stocks, and various vine copulas are employed based on the LMST to model the joint distribution of the selected stocks. This copula-network-based distribution is then used to set the weights of the selected stocks in the portfolio. The empirical investigation involves the component stocks of the S&P 100 index from 2019 to 2023, using a rolling-window framework. The numerical results demonstrate that the proposed method yields satisfactory cumulative returns compared to competitors.

C0438: **Factor multivariate stochastic volatility models of high dimension**
Presenter: **Manabu Asai**, Soka University, Japan
Co-authors: Benjamin Poignard

Building upon the pertinence of the factor decomposition to break the curse of dimensionality inherent to multivariate volatility processes, a factor model-based multivariate stochastic volatility (fMSV) framework is developed that relies on two viewpoints: sparse approximate factor model and sparse factor loading matrix. A two-stage estimation procedure is proposed for the fMSV model: the first stage obtains the estimators of the factor model, and the second stage estimates the MSV part using the estimated common factor variables. The asymptotic properties of the estimators are derived. Simulated experiments are performed to assess the forecasting performances of the covariance matrices. The empirical analysis based on vectors of asset returns illustrates that the forecasting performances of the fMSV models outperform competing conditional covariance models.

CO191 Room BH (S) 2.05 PANEL MODELS Chair: Ralf Wilke

C0793: **A semiparametric panel data model with common factors and spatial dependence**
Presenter: **Juan Manuel Rodriguez-Poo**, Universidad de Cantabria, Spain
Co-authors: Alexandra Soberon, Antonio Musolesi

In the analysis of the Griliches' knowledge capital production function, previous works pointed out the relevance of incorporating slope heterogeneity in the technological parameters, cross-sectional dependence arising simultaneously from common factors and spillovers, and possible nonlinear effects of relevant common observed variables. In order to solve the above problems, a semiparametric model is introduced in a partially linear form that copes simultaneously with all the previous specification issues. The asymptotic properties of the resulting estimators are obtained, and the theoretical findings are further supported for small samples via several Monte Carlo experiments and an empirical application.

C1115: **Binary response dynamic panel data models with switching state dependence**
Presenter: **Eleni Aristodemou**, University of Cyprus, Cyprus

Identification in binary response dynamic panel data models is studied with switching state dependence. Departing from the standard approach of modelling binary response dynamic panel data models, where the last period's choice enters as an additional regressor, switching dependence is allowed where the current period's decision depends on whether this period's choice differs from the last period's choice. This form of correlation causes inertia in individual choices and is suitable for modeling cases where individuals face some form of high "switching costs". This contemporaneous effect in choices, where the choice an individual makes in the current period directly affects the current period's latent utility, results in the model being logically inconsistent, making the model both incomplete and incoherent, which might result in a lack of point identification.

C1163: **On the use of random forests to estimate triangular two-level panel data models with individual fixed effects**
Presenter: **Monika Avila Marquez**, University of Bristol, United Kingdom

The focus is on investigating the use of random forests to estimate triangular simultaneous equation models for panel data with an additive separable individual specific effect and an additive separable disturbance term. First, a model composed of a linear structural equation is considered with one endogenous variable. The endogenous variable presents a nonlinear relationship with the instrumental variables and the exogenous regressors. The parameter of interest is the structural parameter of the endogenous variable. The identification of this parameter is obtained under the assumption of available exclusion restrictions and using a control function approach. The estimation of the parameter of interest is done using two proposed estimators that are composed of two steps. In the first step, the nonlinear reduced form equation is estimated using random forests and the residuals are obtained. In the second step, the residuals are used as an estimated control function for the endogeneity in the structural equation. Later, the functional form assumption is relaxed in the structural equation and a semiparametric structural equation is considered. In this new setting, random forests are used to estimate the nonparametric component of the structural equation. A Monte Carlo simulation is performed to test the performance of the estimators proposed. It is concluded that the estimators perform well, provided that the nuisance parameter can be accurately learnt in the first stage.

C1164: **Fixed effects quantile regression via deconvolutional differencing in short panels**
Presenter: **Martin Mugnier**, Paris School of Economics, France

The focus is on providing point identification results for a quantile regression model with distributional fixed effects. Instead of typical high-level assumptions of nonlinear measurement error models or covariates with dense or large support, a low-level shape restriction is considered: conditional symmetry. Conditional symmetry allows for covariate-heterogeneous quantile effects and arbitrary correlation between the fixed effects and the covariates without ruling out asymmetry of the observed distributions. It is shown how deconvolutional differencing can be applied when at least two measurements are available. Under mild regularity conditions, computationally simple and numerically reliable plug-in estimators are sup-norm consistent and pointwise asymptotically normal as the sample size diverges. Monte Carlo simulations suggest excellent finite-sample performance. The new method is applied to measure the effect of smoking during pregnancy on birth weight.

CO213 Room BH (SE) 2.01 MULTIVARIATE DEPENDENCE STRUCTURES Chair: Zinoviy Landsman

C0276: **Integrating ESG performance in traditional risk measures of mutually dependent risks**
Presenter: **Tomer Shushi**, Ben Gurion University of the Negev, Israel

Recently, socially responsible investments have gained much attention. Environmental, social, and governance (ESG) scores quantify a company's environmental and societal contributions. Traditional risk measures such as expected utility measures and tail-value-at-risk provide different frameworks for measuring risks based on their historical data. The mean-variance measure measures a portfolio risk while capturing the investor's level of risk aversion. A new framework is proposed for incorporating the investor's ESG performance when measuring the risks. The proposed approach provides a way to modify traditional risk measures such that the ESG risk scores are considered. Fundamental features of the proposed approach are explored, which also involves finding the optimal weights from a portfolio containing different stock returns, each having a different ESG score, based on the ESG performance of its company.

C0283: **A minimum variance approach to linear regression with application to actuarial and financial problems**
Presenter: **Zinoviy Landsman**, University of Haifa, Israel
Co-authors: Udi Makov

Uncertainty is intrinsic to statistical, actuarial, and economic models, necessitating accurate quantification for informed decision-making and risk management. A recent study introduces the location of minimum variance squared distance (LVS) risk functional, offering a novel measure of uncertainty. LVS is extended to assess uncertainty in regression models commonly used in actuarial analysis. This functional allows the generation of predictors in the sense of minimum variance squared deviation (MVS). In this context, it is shown that when predicted vector Y follows a symmetric distribution, MVS resembles the traditional minimum expected squared deviation (MES) functional. However, for non-symmetric distributions, MES and MVS exhibit disparities influenced by the joint third moments matrix of distribution P and the covariance matrix of vector Y. The analytical expression is derived for MVS, and a mixed combination of MVS and MES functionals is explored. Two numerical illustrations are provided predicting three components of fire losses: buildings, contents and profits, and predicting returns for six market indices using the returns of their dominant stocks.

C0436: **From colors to numbers: Statistical methods for color theories**
Presenter: **Nicola Loperfido**, University of Urbino, Italy
Co-authors: Cinzia Franceschini, Noemi Loperfido, Marianna Bruto

Colors have been investigated in nearly all fields of human knowledge, including art, history, design, fashion, literature, philosophy, chemistry, physics, marketing, and psychology. Each field of knowledge generated several, often conflicting, color theories. Different field-specific literature on colors shows little attention to each other. Moreover, most theories are poorly grounded in empirical research, and when they are, they use simple and outdated statistical methods. The aim is to address the latter problems through innovative and established multivariate statistical methods, including projection pursuit and invariant coordinate selection. Some preliminary results are illustrated using survey data on a painting of Mondrian, seasonal color analysis and emotional perception of colors in children.

C0452: **Enhancing dependence modeling with compound Archimedean and vine copulas for electricity peak demand estimation**
Presenter: **Moshe Kelner**, University of Haifa and Noga - Israel System Operator, Israel

Copulas serve as an effective and elegant tool for modeling dependence between random variables. However, most copula functions possess only a single dependence parameter, thereby limiting the complexity of the dependence structure they can capture. The Archimedean family of copulas are employed, and the Archimedean inverse generator is modified to enhance the number of dependence parameters while maintaining membership in the family. This enhancement is achieved by compounding the inverse generator with a density function of the dependence parameter. The method is demonstrated using the generalized gamma distribution as a compounding density function for the inverse generator of the dependence parameter of the Clayton copula (exhibiting left tail dependence suitable for the winter peak demand) and the C2-copula (exhibiting right tail dependence suitable for the summer peak demand). This approach extends both copula functions from a single-parameter to a three-parameter function, resulting in the creation of new Archimedean families: The Clayton generalized gamma (CGG) and the Kelner-Landsman-Makov (KLM), respectively. Additionally, the conditional value at risk (VaR) is established and utilized to obtain a confidence interval for one variable given the others. A probability model is proposed for electricity peak demand using these new copula functions. Furthermore, an alternative approach using vine copulas as a means of handling dimensionality is also discussed.

CO330 Room BH (SE) 2.05 SUPERVISED, DEEP, AND REINFORCEMENT LEARNING IN ECONOMIC AND FINANCE Chair: Tato Khundadze

C0486: **Multi-agent deep reinforcement learning and LLM-augmented frameworks for economic policy simulation**
Presenter: **Tohid Atashbar**, IMF, United States

The application of multi-agent deep reinforcement learning (MADRL) and LLM-assisted MADRL is explored in economic policy simulation. It is argued that multi-agent modeling is the most natural approach for analyzing the complex, multi-agent nature of economies, with learning from errors mirroring real-world economic entities' trial-and-error processes. Basic concepts are introduced in MADRL, and modeling considerations are discussed. A simple framework is presented with six agents. The integration of LLM-based decision-making and AI-to-AI communication in MADRL frameworks is also examined. As a proof of concept, an LLM-augmented MADRL framework is showcased, simulating a 4-sector economy with three distinct household preferences. While the approach shows promise for enriching macroeconomic modeling tools, substantial foundational work remains in this nascent field.

C0497: **A nonlinear Gegenbauer process to model the unemployment rate in the G7 countries**
Presenter: **Gilles Dufrenot**, Aix-Marseille School of Economics, France
Co-authors: Ulrich Aiounou

A new model is proposed, combining smooth transition model (STAR) nonlinearities and long memory through Gegenbauer processes. A model is considered where long memory is captured by Gegenbauer polynomials. These processes are widely used in applied econometrics but have never been used in models combining nonlinearity and long memory. Indeed, the long memory property of ARFIMA models is captured, in particular, by exploiting the spectral density in the vicinity of zero. However, long-memory dynamics can also appear if we have long cycles, which requires examining the existence of poles (explosions) of the spectral density, not only at zero but for frequencies close to zero. For an economist modelling the dynamics of the unemployment rate, the difference between ARFIMA and Gegenbauer processes is important in order to differentiate between business cycles (captured by regime-switching models), on the one hand, and long cycles reflecting changes in the structural unemployment rate, on the other. A Lagrange multiplier test and a conditional likelihood-based estimator are proposed. The size and power of the tests are proposed through Monte Carlo simulations.

C0662: **Deep-MacroFin: Informed equilibrium neural network for continuous time economic models**
Presenter: **Goutham Gopalakrishna**, Rotman School of Management, University of Toronto, Canada

A comprehensive framework designed to solve partial differential equations is presented, with a particular focus on models in continuous time economics. This framework leverages deep learning methodologies, including conventional multi-layer perceptrons and the newly developed Kolmogorov-Arnold networks. It is optimized using economic information encapsulated by Hamilton-Jacobi-Bellman equations and coupled algebraic equations. The application of neural networks holds the promise of accurately resolving high-dimensional problems with fewer computational demands and limitations compared to standard numerical methods. This versatile framework can be readily adapted for function approximation, elementary differential equations, and systems of differential equations, even in cases where the solutions may exhibit discontinuities. Moreover, it offers a more straightforward and user-friendly implementation than existing libraries.

C0886: **Modeling cooperative fiscal policy in the Euro area using reinforcement learning and NMPC**
Presenter: **Tato Khundadze**, The New School, United States
Co-authors: Willi Semmler

The purpose is to build on the experience of global shocks, such as the Eurozone crises of 2009-2012 and the economic crises resulting from COVID-19 starting in 2020. It aims to demonstrate the importance of cooperation in terms of monetary and fiscal policies during emergencies. The Euro area is chosen as the sample for testing the models presented in the paper. The case of the Euro area is crucial since its resilience heavily depends on cooperation between different actors within the region. The shocks affecting the nations within the European Union are asymmetric, and the responses to these shocks require coordination, considering the heterogeneous economic structures and levels of economic development. It is built on the previous study on cooperative and fiscal policies in the Euro Area, which uses nonlinear model predictive control (NMPC) to trace the path of key macroeconomic variables (inflation rate, interest rate, output gap, government gross debt, and price level) dynamics under non-cooperative and cooperative scenarios. The prior study is further extended by applying reinforcement learning to model the dynamics of the mentioned variables in a multi-agent environment.

CO219 Room BH (SE) 2.10 CWS SESSION: STATISTICS AND DATA SCIENCE FROM WOMEN AROUND THE GLOBE Chair: Cynthia Bland

C0323: **Modelling drought classes time series for groundwater drought assessment and prediction in Algarve region**
Presenter: **Elsa Moreira**, NOVA School of Science and Technology (NOVA FCT), Portugal
Co-authors: Maria Neves

Log-linear quasi-association models have been successfully applied to analyze and predict drought class transitions derived from standardized precipitation index (SPI) time series in Portugal. This kind of model proved to be suitable for fitting the SPI drought transitions and is considered a reliable tool for capturing the dynamics of drought severity changes since it models the probabilities associated with transitions in drought severity over specific time periods. In the context of groundwater drought monitoring, the standardized groundwater index (SGI) is used and is computed from groundwater levels available from the SNIRH piezometric network. The aim is to employ similar models to model the transitions between SGI drought classes and use them to analyze and predict transitions in groundwater drought classes one or two months in advance. The purpose is also to evaluate the effectiveness of these tools in predicting short-term transitions in groundwater drought. The findings contribute to improving water management practices and enhancing early warning systems to mitigate the impacts of drought in the Algarve, with potential applications in other parts of the world.

C0471: **GMANOVA modelling for volatile data**
Presenter: **Sayantee Jana**, Indian Institute of Technology Hyderabad, India

Generalized multivariate analysis of variance (GMANOVA) models are linear models useful for the analysis of longitudinal data, which are repeated measurements of a continuous variable from several individuals across any ordered variable such as time, temperature, pressure, etc. GMANOVA models are widely used in economics, social sciences and medical research. However, despite financial data being time-varying, the traditional GMANOVA model has limited to no applications in finance due to the volatile nature of such data. This, in turn, makes financial data the right candidate for Multivariate t (MT) distribution, as it allows for outliers in the data to be modelled due to its heavy tails. In fact, portfolio analysis, including mutual funds and capital asset pricing, are all modelled using elliptical distributions, especially MT distribution. The classical GMANOVA model assumes multivariate normality, and hence, the inferential tools developed for the classical GMANOVA model may not be appropriate for heavy-tailed data. The sensitivity of inferential tools developed under multivariate normality under volatile data is first explored, and then inferential tools are developed for the GMANOVA model under the MT distribution. The practical implementability of the proposed method is demonstrated on a financial dataset.

C0792: **A metric based on the efficient determination criterion**
Presenter: **Veronica Andrea Gonzalez-Lopez**, University of Campinas, Brazil

The concept of metrics based on the Bayesian information criterion (BIC) is extended to achieve a strongly consistent estimation of partition Markov models (PMMs). A set of metrics is introduced, drawn from the family of model selection criteria known as efficient determination criteria (EDC). This generalization extends the range of options available in BIC for penalizing the number of model parameters. The relationship is formally specified and determines how EDC works when selecting a model based on a threshold associated with the metric. Furthermore, the penalty options are improved within EDC, identifying the penalty ln(ln(n)) as a viable choice that maintains the strongly consistent estimation of a

PMM. To demonstrate the utility of these new metrics, those are applied to the modeling of three DNA sequences of dengue virus type 3, endemic in Brazil in 2023.

C0445: **Adapting student performance evaluation in the AI era**
Presenter: **Suhwon Lee**, University of Missouri, United States

The integration of artificial intelligence (AI) into statistical education is changing learning methodologies and offering transformative opportunities for both educators and students. AI's impact on statistical education is evident in its ability to provide tailored learning experiences. Machine learning algorithms and intelligent tutoring systems offer personalized feedback and assessments, identifying areas of difficulty for individual students. AI-powered statistical software assists students in handling complex datasets, enabling them to focus on conceptual understanding rather than computational intricacies. Collaborative learning experiences are facilitated through AI-driven platforms equipped with natural language processing capabilities. Despite these advancements, ethics should be considered when adopting AI in statistical education. Addressing concerns related to data privacy, algorithmic bias, and the risk of overreliance on AI tools is crucial for responsible implementation. Additionally, continuous professional development for educators is essential to maximize the effective integration of AI technologies into teaching practices. The purpose is to explore the key aspects of AI adoption in education, focusing on personalized learning experience and collaborative platforms.

CO024 Room BH (SE) 2.12 RECENT ADVANCES IN SPATIAL AND SPATIO-TEMPORAL DATA MODELING Chair: Rajarshi Guhaniyogi

C1161: **STimage-1K4M: A large scale dataset for spatial transcriptomics**
Presenter: **Didong Li**, University of North Carolina at Chapel Hill, United States

Recent advances in multi-modal algorithms have driven and been driven by the increasing availability of large image-text datasets, leading to significant strides in various fields, including computational pathology. However, in most existing medical image-text datasets, the text typically provides high-level summaries that may not sufficiently describe sub-tile regions within a large pathology image. For example, an image might cover an extensive tissue area containing cancerous and healthy regions, but the accompanying text might only specify that this image is a cancer slide, lacking the nuanced details needed for in-depth analysis. The aim is to introduce STimage-1K4M, a novel dataset designed to bridge this gap by providing genomic features for sub-tile images. STimage-1K4M contains 1,149 images derived from spatial transcriptomics data, which captures gene expression information at the level of individual spatial spots within a pathology image. Specifically, each image in the dataset is broken down into smaller sub-image tiles, with each tile paired with 15,000-30,000 dimensional gene expressions. With 4,293,195 pairs of sub-tile images and gene expressions, STimage-1K4M offers unprecedented granularity, paving the way for a wide range of advanced research in multi-modal data analysis and innovative applications in computational pathology and beyond.

C1169: **Nonstationary elastic space-time (NEST) Kriging and solar irradiance studies**
Presenter: **William Kleiber**, University of Colorado, United States

As the power grid moves to a more renewable future, energy sources from weather-driven phenomena such as solar power will form an increasingly large portion of electricity generation. The variability, non-Gaussianity and intermittency of solar resources challenge current grid operation paradigms, and realistic data scenarios are required for grid planning and operational studies. However, such data are not available at the space-time resolution needed for realistic grid models. Given sparse spatial samples, a framework for spatiotemporal prediction in a functional data analysis framework is introduced when data exhibit nonstationary phase misalignment. The approach is illustrated on a challenging high-frequency irradiance dataset and compared with existing methods.

C1176: **Bayesian geostatistics using predictive stacking**
Presenter: **Lu Zhang**, University of Southern California, United States

Bayesian predictive stacking is presented for geostatistical models, where the primary inferential objective is to provide inference on the latent spatial random field and conduct spatial predictions at arbitrary locations. Analytically tractable posterior distributions are exploited for regression coefficients of predictors and the realizations of the spatial process conditional upon process parameters. Such inference is subsequently combined by stacking these models across the range of values of the hyper-parameters. Stacking of means and posterior densities are devised in a manner that is computationally efficient without resorting to iterative algorithms such as Markov chain Monte Carlo (MCMC) and can exploit the benefits of parallel computations. Novel theoretical insights are offered into the resulting inference within an infill asymptotic paradigm and through empirical results showing that stacked inference is comparable to full sampling-based Bayesian inference at a significantly lower computational cost.

C1319: **Markov random fields with proximity constraints for spatial data**
Presenter: **Sudipto Saha**, Florida State University, United States
Co-authors: Jonathan Bradley

The conditional autoregressive (CAR) model, simultaneous autoregressive (SAR) model, and its variants have become the predominant strategies for modeling regional or areal-referenced spatial data. The overwhelming wide-use of the CAR/SAR model motivates the need for new classes of models for areal-referenced data. Thus, we develop a novel class of Markov random fields based on truncating the full-conditional distribution. We define this truncation in two ways leading to versions of what we call the truncated autoregressive (TAR) model. First, we truncate the full conditional distribution so that a response at one location is close to the average of its neighbors. This strategy establishes relationships between TAR and CAR. Second, we truncate the joint distribution of the data process in a similar way. This specification leads to connections between TAR, SAR, and the nearest-neighbor Gaussian process (NNGP) model. Our Bayesian implementation does not use Markov chain Monte Carlo (MCMC) for Bayesian computation, and generates samples directly from the posterior distribution. Moreover, TAR does not have a range parameter that arises in the CAR/SAR models, which can be difficult to learn. We present the results of the proposed truncated autoregressive model on several simulated datasets and on a dataset of average property prices.

CC510 Room K0.19 STATISTICAL METHODS AND APPLICATIONS II Chair: Kalliopi Mylona

C1733: **The pvars R-Package: VAR modeling for heterogeneous panels**
Presenter: **Lennart Empting**, Uni Goettingen, Germany

pvars offers a seamless implementation of vector autoregressive (VAR) methods for heterogeneous panel data. The R-package comprises panel cointegration rank tests, which can account for cross-sectional dependence and structural breaks in the deterministic term. The implemented panel SVAR models can be estimated under these specifications with pooled cointegrating vectors and identified by various panel identification procedures. We review these methods and present their modular implementation in R. The empirical illustrations replicate examples from the literature step-by-step and guide the pvars' users into conducting their own analyses.

C1735: **Spatial causal analysis: Case study of testicular cancer and PFAS exposure in Veneto, Italy**
Presenter: **Allegra Sartore**, University of Padova, Italy
Co-authors: Dolores Catelan, Tony Fletcher, Cristina Canova, Mirko Berti, Giorgia Stoppa, Annibale Biggeri

Causal inference is essential in epidemiologic research and public health. Ecological studies using spatial or spatio-temporal data face challenges with spatial autocorrelation, violating the "no interference" assumption. An ecological regression study analyzed orchiectomy rates for testicular cancer across 21 municipalities in Veneto, Italy, where water was contaminated by PFAS. Exposure data included municipality-specific geometric means of PFOA serum concentrations (GMPFOA), adjusted for individual covariates. A Bayesian hierarchical spatial model was specified with a propensity score model on GMPFOA and socio-demographic variables, plus a shared spatial random component. The disease model included GMPFOA, socio-demographic variables, the shared spatial component, and a specific random component for response spatial patterns. The regression coefficient for the exposure of interest (logRR per 30 ng/L PFOA) was 0.315 (95% Credibility Interval: -0.018, 0.668). Using the propensity score approach, an adjusted causal regression coefficient of 0.829 (95% CrI: 0.078, 1.616) was found. This approach incorporates two spatial random components to meet assumptions for causal interpretation of the regression coefficients. Sensitivity analysis discusses structural equation models and instrumental variables. A dose-response association between testicular cancer risk and PFOA exposure is confirmed.

C1734: **Latent space models for grouped multiplex networks**
Presenter: **Alexander Kagan**, University of Michigan, United States
Co-authors: Ji Zhu, Liza Levina

Latent space models such as the stochastic block model and the random dot product graph are popular ways of modeling single-layer networks. However, their application to more complex network structures has not received a lot of attention so far. MacDonald et al. [2021] made a substantial step in this direction by introducing the MultiNeSS model allowing to extract a latent space component shared by a sample of multiplex networks: multiple, heterogeneous networks observed on a shared node set together. This work, however, has an apparent limitation arising from the fact that groups of networks within this sample may have individual group structures besides the one that is common for the whole sample. Such group stratification may arise when for each network in a sample we additionally observe a categorical attribute, e.g.together with the patients' brain region connectivity networks we can have access to their gender, ethnicity, age group, or control/treatment label. For this more general model that we call GroupMultiNeSS, we establish identifiability, develop a fitting procedure using convex optimization in combination with a nuclear norm penalty, and prove a guarantee of recovery for the latent positions. We compare the model with the original MultiNeSS model in various synthetic and real-world scenarios and observe an apparent improvement in the modeling accuracy when the group component is accounted for.

CC438 Room K2.40 MACHINE LEARNING IN APPLICATIONS Chair: Abdelaati Daouia

C0194: **Can higher data frequency lead to more accurate stock market predictions: Nasdaq 100 and DAX cases**
Presenter: **Nuno Ferreira**, ISCTE-IUL, Portugal
Co-authors: Diana Aldea Mendes, Vivaldo Mendes

The aim is to assess if the frequency of time series is associated with increased forecast accuracy. Two different time series from the G7 countries, the NASDAQ100 and the DAX, are examined for a period of five minutes, as well as daily frequency. The employed algorithms are deep learning recurrent neural networks that are particularly suited for a variety of variations of Long Short-Term Memory (LSTM) structures (LSTM, BiLSTM). A random search over the hyperparameters was employed to determine the architecture that minimizes the loss function. A better outcome is obtained for the 5-minute daily frequency for both datasets, with the forecast increased by 1%.

C1055: **Enhancing sign language translation with real-time AI technology**
Presenter: **Elisa Cabana G. del Vall**, CUNEF, SL, Spain

A primary challenge for the deaf and hearing-impaired community stems from the communication gap with the hearing society, which can greatly impact their daily lives and result in social exclusion. To foster inclusivity in society, the endeavor focuses on developing a cost-effective, resource-efficient, and open technology based on artificial intelligence designed to assist people in learning and using sign language for communication. Specifically, the aim is to create a computer vision system for American sign language (ASL) fingerspelling classification in real time that can serve as a learning application. For this purpose, an extensive dataset of images of ASL alphabet signs has been compiled. Several neural network classification models are compared and implemented into the final real-time system with the overall best performance metrics. Valuable insights for the sign language translation scenario and significant advances in ongoing academic research are provided.

C1342: **EEMD-ELN regression for multi-scale relationships: Application for rainfall prediction**
Presenter: **Ahmed Alsayed**, University of Bergamo, Italy

The increase in extreme weather events, including unseasonal rainfall and floods due to climate change, encourages more accurate, timely rainfall forecasts. In general, raw environmental data are often multi-scale, non-stationary, and highly intercorrelated, leading to poor prediction accuracy and reliability. To deal with these gaps, a hybrid approach is proposed using the ensemble empirical mode decomposition (EEMD) combined with elastic net (ELN) penalized regression to overcome these challenges. The proposed method presents several advantages; firstly, the original predictors of rainfall time-series data are decomposed using EEMD into intrinsic mode functions (IMFs) and one residual component. Secondly, this approach detects the relationship between a response variable and the new predictors at different time scales. The performance of the proposed method is proven by using atmospheric variables for the city of Basel in Switzerland over the period from 1/1/2022 to 31/12/2023. The main finding demonstrated that the proposed ELN-EEMD model outperformed the other models. Moreover, the final estimated model shows that rainfall is negatively affected by the high degree of temperature and extreme waves represented by the high-frequency IMF while it is positively affected by wind gusts and humidity at various high degrees.

C1588: **Regression trees for analyzing longitudinal health data streams: A comparative study**
Presenter: **Ines Sousa**, Minho University, Portugal

Chronic kidney disease (CKD) is characterized by kidney damage or an estimated glomerular filtration rate (eGFR) of less than 60 ml/min per 1.73 square meters for three months or more. The performance of six tree-based machine learning models - Decision Trees, Random Forests, Bagging, Boosting, Very Fast Decision Tree (VFDT), and Concept-adapting Very Fast Decision Tree (CVFDT)- are evaluated on longitudinal health data. Longitudinal data, where individuals are measured repeatedly over time, provide an opportunity to predict future trajectories using dynamic predictions that incorporate the entire historical dataset. These predictions are essential for real-time decision-making processes in healthcare. The dataset comprised 406 kidney transplant patients, spanning from January 21, 1983, to August 16, 2000. It captures 120 time points over the first 119 days post-transplant, including baseline glomerular filtration rates (GFR), along with three static variables: weight, age, and gender. Data preprocessing involved robust imputation techniques to handle missing data, ensuring consistency and trend accuracy. The models were trained to predict health outcomes starting from the eight-day post-transplant, progressively incorporating daily values to predict subsequent days up to day 119. Model performance was evaluated using mean squared error (MSE) and mean absolute error (MAE) through data partitioning and cross-validation techniques.

CC416 Room S0.03 STATISTICAL MODELLING Chair: Jan Gertheiss

C0329: **A dynamic extension of the Massey's rating system with an application in basketball**
Presenter: **Paolo Vidoni**, University of Udine, Italy

A flexible, dynamic extension of the popular Massey's method is proposed for rating players and teams involved in sports competitions. The original Massey's approach is static since the computation of a team rating is based on the strength of the opponent teams evaluated at the current time. The proposed dynamic extension updates the team rating, considering the strength of the opponent teams evaluated at the time when the matches were played. This approach adequately takes into account the fact that teams' capabilities change over time. In addition, the associated statistical model is flexible and easily extensible to improve its predictive ability. The proposed method accounts for the evolution of both the offensive and the defensive rating for each team, instead of a single general rating. An application of the new rating procedure to the Euro league Basketball 2018-2019 is presented.

C1607: **On the choice of parametric link for binomial-response generalized linear models**
Presenter: **Takis Besbeas**, Athens University of Economics and Business, Greece
Co-authors: Dimitrios Stroungis

Logistic and probit regression are the two most commonly used techniques to analyse binary and binomial response data. Although these models have desirable properties, several authors have noted their sensitivity to outliers and have recognised that the use of a larger class of link functions within a binomial generalised linear model may have advantages in practice. It is illustrated that there is considerable choice beyond the use of the logit and probit, resulting from the use of a range of cumulative distribution functions to relate the probability of success to the linear predictor. Extending work on binary outcomes, a Gosset regression model is considered based on the CDF of a t-distribution with a free degree of freedom parameter, which includes the logit, probit and Cauchit models as special cases. The framework allows the choice of link function to be addressed through the estimation of the degrees of freedom parameter. Estimation of that parameter is described by profile likelihood, and the performance of the method is evaluated using simulation, illustrating that estimation is uncertain in the case of binary outcomes but improves with the number of trials. The impact of misspecification of the link function on the parameter estimates is evaluated. Results demonstrate that it is prudent to convey a more skeptical attitude about the conventional choice of logit and probit links in practice, especially when the number of trials is large.

C1633: **On some bivariate transmuted distribution models**
Presenter: **Violetta Piperigou**, University of Patras, Greece

Recently, various transmuted probability models have appeared in the literature by transforming a base distribution into its extended counterpart and in some particular models, a theoretical interpretation has been obtained. Bivariate transmuted models are proposed which arise from mixtures of joint order statistics and from certain generalized (stopped-sum) distribution models. Some examples are discussed in detail, and their properties are examined.

C1487: **A cognitive diagnostic model for matching format tests**
Presenter: **Rinhi Higashiguchi**, The University of Tokyo, Japan
Co-authors: Kentaro Fukushima, Kensuke Okada

Matching format tests are widely used in achievement assessments and psychological evaluations, where a respondent is presented with a list of test items and response alternatives and asked to match each response alternative with a test item. However, the development of cognitive diagnostic models (CDMs) tailored to matching format tests remains unexplored. CDMs are models designed to identify the presence or absence of multiple fine-grained attributes, enabling more precise and informative assessments. When matching format tests are analyzed using conventional CDMs, there is a risk of overestimating respondent parameters due to the unique structure of the test format. To address this issue, a deterministic inputs are proposed, noisy, and gate (DINA) model for matching test formats. Amongst CDMs, the DINA model assumes that all attributes are required to correctly answer an item (in exam settings) and is parsimonious and easily interpretable. A comparison between the conventional DINA model and the proposed model was conducted through a simulation study. The results suggest that the proposed DINA model for matching format items offers a promising approach to improving the diagnostic accuracy of CDMs when applied to this unique test format.

CC458 Room BH (SE) 2.09 MACROECONOMETRICS Chair: Masayuki Hirukawa

C0519: **The macroeconomic effects of internal promotions**
Presenter: **Demetris Koursaros**, Cyprus University of Technology, Cyprus

The purpose is to explore the macroeconomic implications of a hierarchical payment structure according to which a worker's wage within a firm depends on his job rank. A model of the labor market in which some firms offer "career" jobs is developed. Workers in career jobs start at a low rank and can get a wage increase through promotion to a higher rank. It is shown that it is optimal for career firms to incentivize their employees by increasing the wage spread between low- and high-rank positions. This way, they can elicit more effort from the mass of the workers in low ranks while rewarding handsomely only the very few that get promoted. As workers in low ranks are compensated for their higher effort through the option value of a promotion, the model departs from the standard assumption that wages increase with effort or outside option. It is shown that this payment structure, which has been largely overlooked in the macroeconomic literature, can provide interesting insights into various puzzles, such as the cyclicality of the labor wedge and the gender wage gap.

C1543: **Noisy past and business cycles**
Presenter: **Naoko Hara**, Seikei University, Japan

While standard macroeconomic models assume perfect information about the current and past states of the economy in real-time, accurate information often becomes available only after a long delay in practice. To assess the impacts of mismeasurement on economic activities and expectations, a structural vector autoregression model is estimated using different data vintages of the U.S. productivity over 1968Q1-2008Q1. The estimation results show that the measurement errors in early releases of productivity data significantly affect the underlying economic fundamentals and expectations in the short run. Furthermore, professional forecasts from the survey of professional forecasters positively respond to these errors, particularly to long-lived measurement errors that persist even after multiple data revisions. The heavily revised data may not yet fully reflect recent shifts in industrial structures and other key information that describe recent structural changes because the information is only incorporated into the data through comprehensive revisions generally conducted every five years. The long absence of such information in the data could explain why the long-lived errors continue to impact expectations. The findings suggest that this absence of information can influence the economy through expectations.

C0180: **A high-dimensional GDP-at-risk and inflation-at-risk for the Euro area**
Presenter: **Matteo Santi**, Bank of Italy, Italy

Adapted from tools originally developed in the financial risk management literature, the GDP-at-risk and the inflation-at-risk are standard measures of tail risk in modern macroeconometrics. These indicators are estimated for the Euro Area and its member states by leveraging a high-dimensional dataset in the construction of time-varying conditional distributions of GDP growth and inflation. The distributions obtained at the country level are used to assess how the synchrony of the EA countries' business cycles has evolved since the introduction of the Euro. Results indicate significant asymmetries in the balance of upside and downside risks for both GDP and inflation, and a persistently weak synchrony of the left tails of the GDP growth distributions during episodes of crisis.

C1636: **A test of exogeneity in structural vector autoregressions with external instruments**
Presenter: **Luca Fanelli**, University of Bologna, Italy
Co-authors: Giovanni Angelini, Giuseppe Cavaliere

A novel test is introduced for instrument exogeneity that is robust to proxy strength and does not require auxiliary information beyond SVAR restrictions and the instruments themselves. The test uses a Waldt-type logic to assess whether specific moment conditions implied by the proxy-SVAR are consistent with the hypothesis of no contamination of the instruments by the non-target shocks. To construct the test statistic, a consistent estimator of key elasticity parameters is required, independent of the instruments. This estimator is derived from the dynamic structure of the SVAR, under a set of overidentifying restrictions similar to ‘Byron-type’ constraints used in simultaneous equation systems. The test is reliable only if the Byron-type restrictions hold within the SVAR. Conditional to the non rejection of the Bayron’s restrictions, the asymptotic inference on instrument exogeneity based on this estimator is valid. Extensive Monte Carlo simulations are conducted to evaluate the finite-sample performance of the test. Its practical utility is demonstrated by revisiting some empirically prominent proxy-SVARs from the existing literature.

Sunday 15.12.2024 15:50 - 17:30 Parallel Session I – CFECMStatistics2024

CI053 Room Auditorium CHALLENGES IN PREDICTION IN TIMES OF CLIMATE RISKS AND GLOBAL CONFLICTS Chair: Willi Semmler

C0153: **Changes in precipitation: Trends, seasons, breaks, and memory**
Presenter: **Harry Haupt**, University of Passau, Germany

Understanding precipitation processes is crucial for grasping the changing global and local precipitation patterns noted in recent IPCC reports. Precipitation variability and uncertainty are expected to rise by the mid-21st century, continuing trends in regions like the northern high latitudes and tropical land areas. Projections indicate increased mean precipitation for Asia and Polar regions, decreases for Africa and Australia, and region-specific changes in Europe and the Americas. The rise in heavy precipitation and aridity globally poses significant challenges for ecosystems, society, and water management. Detecting and attributing changes to human influences requires careful analysis of evolving climate conditions and their interaction with natural variability. The memory parameters of global precipitation over the last 60 years are investigated, taking into account level shifts, trend and seasonal patterns, and low-frequency contamination. Considerable heterogeneity is found in both the components and the memory properties of the monthly precipitation anomalies. While long-range dependence is a stylized fact of precipitation processes, it is subject to considerable uncertainty and exhibits substantial variability over specific time spans and at certain climatological and geographic locations.

C0154: **Statistical contributions in conflict research**
Presenter: **Goeran Kauermann**, LMU Munich, Germany

The field of statistical conflict research is showcased. Armed conflicts are considered in regions of Africa with at least one fatality. The reliable data provided by the Peace Research Institute Oslo (PRIO) is used as a data source. Firstly, the way in which statistical models can be used to predict future conflicts is demonstrated. A hurdle model is applied, which allows the proper incorporation of quantities relevant from the political science perspective. Secondly, Syria is considered and explored in terms of how remote sensing data from satellites can improve prediction accuracy. In particular, machine learning tools are included, and those are compared with statistical approaches. Finally, the spread of conflicts is looked at in time and space, which is often omitted in purely machine learning approaches.

C0155: **Climate risks and multiple objectives decision-making: Model-guided and empirical assessments**
Presenter: **Willi Semmler**, New School for Social Research, United States

Climate change and climate risks are studied in the context of long-run growth models In model variants with externalities, they introduce new policy components such as carbon emissions, abatement efforts, and disruptive extreme events through temperature rise. There are often conflicting goals defined, for example, between growth and climate risk control. In short- and medium-run macro models, however, new objectives such as CO2 emission, have been introduced. Yet, for macroeconomic policies, it holds that one faces many conflicting objectives, for example, between inflation rate, labour market performance, and financial instability. As growth-oriented and macro-oriented tools, instruments, and policies have proven useful, yet often, the multiple (conflicting) goals appear to reduce the effectiveness of the policy tools. We introduce the method of multiple objectives control in higher-dimensional climate-economy complex systems that allow treatment of both multiple objectives in growth- and macro-oriented models. An assessment of the effectiveness of multiple objectives control and a suggestion for providing prioritization and weights on different objectives will be given. The numerical procedure is called Linear Scalarization. It allows for dynamically providing a Pareto front and finding the weights for re-balancing imbalances. A VECM with stationary and non-stationary variables is used. This also allows us to assess what weights certain objectives should have.

CO015 Room S-2.23 ANOMALIES IN ASSET PRICING (VIRTUAL) Chair: Nathan Lassance

C0167: **Limits to arbitrage to explain portfolio gains from asset mispricing**
Presenter: **Nathan Lassance**, UCLouvain, Belgium
Co-authors: Alberto Martin-Utrera

The efficiency gains from asset mispricing are studied through the lenses of an optimal arbitrage portfolio exploiting a large set of firm characteristics. In particular, the tangency portfolio is decomposed into a level component that captures time-series return variation and an optimal arbitrage component that exploits mispricing. It is found that the arbitrage portfolio offers considerable efficiency gains. However, these gains only survive arbitrage impediments such as estimation risk, transaction costs, and short-sale constraints during high-sentiment periods. This finding are capitalized on to construct arbitrage-friendly portfolios that effectively utilize asset mispricing to span better the achievable efficient frontier.

C0182: **The expected returns on machine-learning strategies**
Presenter: **Mihail Velikov**, Pennsylvania State University, United States
Co-authors: Vitor Azevedo, Christopher Hoegner

The expected returns of machine learning-based anomaly trading strategies are assessed, accounting for transaction costs, post-publication decay, and the post-decimalization era of high liquidity. Contrary to claims in prior literature, more sophisticated machine learning strategies are profitable, earning net out-of-sample monthly returns of up to 1.42%, despite having turnover rates exceeding 50% and selecting some difficult-to-arbitrage stocks. A trading strategy that employs a long short-term memory model to combine anomaly characteristics yields a six-factor generalized (net) alpha of 1.20% (t-stat of 3.46). While prevalent cost-mitigation techniques reduce turnover and costs, they do not improve net anomaly performance. Overall, return predictability is documented from deep-learning models that cannot be explained by common risk factors or limits to arbitrage.

C0500: **Facts, momentum and factor momentum**
Presenter: **Pedro Barroso**, Universidade Catolica Portuguesa, Portugal
Co-authors: Haoxu Wang

Factor momentum recently joined the ongoing debate over the causes of stock momentum. According to this explanation, momentum in well-known off-the-shelf factors - or in principal component factors responsible for large commonalities in stock returns - greatly subsumes momentum in individual stocks. It is found that neither form of factor momentum can explain any previously proposed momentum driver; conversely, all other drivers combined can subsume both forms of factor momentum. Also, compared to previous drivers, factor momentum does not exhibit superior performance in capturing other momentum-like anomalies. Like the competing models, it cannot explain stock momentum conditionally. Moreover, it cannot explain stock momentum after accounting for transaction costs, while these can explain the persistence of factor momentum, especially in less systematic factors.

C1118: **Why do asset pricing models fail in equilibrium**
Presenter: **Alejandro Lopez Lira**, University of Florida, United States

The equilibrium limitations of asset pricing models are examined with rational, risk-averse investors who have incomplete information about expected returns and covariances. In equilibrium, the Sharpe ratio of the achievable optimal portfolio is bounded, including those implied by asset pricing models. In contrast, the Sharpe ratio of the optimal portfolio under the true data-generating process grows with the difficulty of the estimation problem. Mispricing is related to their squared difference and increases as the estimation problem becomes more challenging. The model makes several novel predictions: (1) pricing errors of parsimonious factor models increase with the number of assets or state variables, even without uncertainty about expected returns, hence diminishing the cross-sectional correlation between betas and average returns; (2) anomalies are expected to proliferate in high dimensions; and (3) multiple high Sharpe ratio strategies can coexist without subsuming each other. Strong empirical support is found for these predictions.

CO332 Room S-2.25 INTEGRATIVE ANALYSIS OF MULTI-SOURCE AND MULTI-WAY DATA Chair: Thierry Chekouo

C0772: A non-iterative algorithm for structural equations modeling

Presenter: **Arthur Tenenhaus**, Laboratoire Signaux et Systemes, France
Co-authors: Michel Tenenhaus

Structural equation modeling is presented with factors and composites in the framework of the basic design. Two estimation methods are discussed. The first one is a new non-iterative method based on singular value decomposition calculations (SVD-SEM). This new approach produces consistent and asymptotically normal estimators of the parameters. Then, the restricted maximum-likelihood approach (ML-SEM) is described. SVD-SEM and ML-SEM are compared in a Monte Carlo simulation on a nonrecursive model with factors and composites and in a case study on a model with factors.

C0840: Partially characterized topology guides reliable anchor-free scRNA-integration

Presenter: **Leying Guan**, Yale University, United States

Single-cell RNA sequencing (scRNA-seq) is increasingly applied to obtain biological insights at cellular resolution, with scRNA-seq batch integration a key step before downstream statistical analysis. Despite the plethora of methods proposed, achieving reliable batch correction while preserving the heterogeneity of biological signals that define cell type continues to pose a challenge, with existing methods' performance varying significantly across different scenarios and datasets. ScCRAFT, a deep-learning model designed to segregate cell-state-related biological signals from batch effects for reliable multi-batch scRNA-seq integration, is proposed. ScCRAFT comprises three main components: an autoencoder that targets the extraction of biological signals, a multi-domain adaptation loss aimed at eliminating batch effects and an innovative dual-resolution triplet loss component for preserving topology within each batch, which is introduced as an effective mechanism to counteract the over-correction effect of domain adaptation loss amid heterogeneous cell distributions across batches. It is shown that scCRAFT effectively manages unbalanced batches, rare cell types, and batch-specific cell phenotypes in simulations and surpasses state-of-the-art methods in a diverse set of real datasets.

C0862: Latent variable methods for multi-view high-dimensional data

Presenter: **Katrijn Van Deun**, Tilburg University, Netherlands

Research in many disciplines relies more and more on intensive collections of data representing several points of view. For example, in studying obesity or depression as the outcome of environmental and genetic influences, researchers increasingly collect survey, dietary, biomarker and genetic data from the same individuals. Revealing the variables that are linked throughout these different types of data gives crucial insight into the complex interplay between the multiple factors that determine human behaviour, e.g., the concerted action of genes and environment in the emergence of obesity or depression. Although linked high-dimensional multiview data form an extremely rich resource for research, extracting meaningful and integrated information is challenging and not appropriately addressed by current statistical methods. The challenge is to select those variables that are linked throughout the different blocks, and this eludes currently available methods for data analysis. The first problem is that relevant information is hidden in a bulk of irrelevant variables with a high risk of finding incidental associations. Second, the sources are often very heterogeneous, which may obscure apparent links between the shared mechanisms. The challenges associated with the analysis of large-scale multiview data are discussed, and a sparse common and distinctive latent variable approach is presented to address the challenges.

C0994: Multi-view multivariate mediation analysis

Presenter: **Sandra Safo**, University of Minnesota, United States

Many biomedical studies generate data from multiple sources or views with the main goal of integrating these diverse but complementary data for deeper biological insights. Most existing integrative analysis methods only consider associations among the views and an outcome without inferring potential causal relationships. Mediation analysis explores causal relationships between exposures and an outcome by including a mediator as an intermediate variable. Existing mediation analysis methods consider only single variate and single view exposures, and none incorporate multi-view exposures. Multi-view multivariate mediation analysis (MMM) is proposed, which considers both high-dimensional multivariate exposures and mediators and incorporates multi-view exposures. MMM integrates multi-view exposures by identifying disentangled common drivers accounting for indirect effects via a multivariate mediator and direct effects to be estimated separately. Simulation studies are used to demonstrate the effectiveness of MMM in comparison with other methods. MMM is applied to data from the ADNI study to explore the potential causal relationship between multiomics exposures and Alzheimer's Disease progression via genetic mediators.

CO162 Room S-1.01 HITEC: TOPICS IN FINANCIAL ECONOMETRICS Chair: Alessandra Amendola

C0494: On periods of extreme asset price volatility to signal the beginning of a recession

Presenter: **Fabio Spagnolo**, The University of Messina, Italy

The interrelationship between financial markets and real economic activity is investigated. A procedure is proposed for analyzing links between stock market volatility and output growth based on a bivariate Markov switching model. The method provides a convenient way of analyzing the predictive content of different series first and second moments. An empirical application of this procedure to the U.S. is examined and discussed.

C0690: Quantifying uncertainty: A new era of measurement through large language models

Presenter: **Simon Stalder**, University of Lugano, Switzerland
Co-authors: Francesco Audrino, Jessica Gentner

The aim is to present an innovative method for measuring uncertainty using large language models (LLMs), offering enhanced precision and contextual sensitivity compared to the conventional methods used to construct prominent uncertainty indices. By analyzing newspaper texts with state-of-the-art LLMs, the approach captures nuances often missed by conventional methods. Indices are developed for various types of uncertainty, including geopolitical risk, economic policy, monetary policy, and financial market uncertainty. Findings show that shocks to these LLM-based indices exhibit stronger associations with macroeconomic variables, shifts in investor behavior, and asset return variations than conventional indices, underscoring their potential for more accurately reflecting uncertainty.

C0988: Does sustainability impact tail risk measurement: Evidence from a novel text-based ESG indicator

Presenter: **Alessandra Amendola**, University of Salerno, Italy
Co-authors: Vincenzo Candila, Peter Winker, Shahram Dehghan Jabarabadi

With the increasing trend of public interest, literature, and legislation in environmental, social, and governance (ESG) and the lack of statistical indicators, the paper proposes a novel daily ESG indicator that uses scrapped news articles and textual analysis. The news article source is selected in a way that ensures the consistency of data over time and across various contexts and provides sufficient observations during the investigated period (from 2004 to 2024). To validate the proposed indicator, graphical approaches and tail risk measurements are employed. The evaluations reveal the outperformance of tail risk measurement models with the ESG indicator as an exogenous variable. The presentation of a daily ESG indicator offers a reliable, long-term time series independent of specific corporate influences to enrich risk management and portfolio optimization strategies by quantifying ESG impact and crafting data-driven ESG policies.

C1142: **(Quantile) spillover indexes: Simulation-based evidence, confidence intervals and a decomposition**
Presenter: **Giovanni Bonaccolto**, University of Enna Kore, Italy
Co-authors: Massimiliano Caporin, Jawad Shahzad

Quantile-spillover indexes have recently become popular for analyzing tail interdependence. It is shown that the estimation of spillover indexes is affected by a positive distortion when the parameters of the underlying fitted models are not evaluated with respect to their statistical significance or are not estimated subject to regularization. The distortion is reduced for increasing sample sizes, thanks to the consistency of estimators, or by filtering out non-significant parameters, even if in small samples, it does not fully disappear due to type I error. In the next step, a simulation-based approach is introduced to recovering confidence intervals from quantile spillover indexes. In addition, an algebraic decomposition of quantile spillover is put forward, separating the dynamic interdependence from the contemporaneous interdependence (due to residual correlation). Empirical evidence on equity sector indices shows that distortions on real data are sizable, and the decomposition points out that most of the spillover is due to contemporaneous effects. All of the results extend and are confirmed for the Spillover index.

CO254 Room S-1.04 ESTIMATION FOR JUMP PROCESS Chair: Marie du Roy de Chaumaray

C0555: **Estimation of a pure-jump stable Cox-Ingersoll-Ross process**
Presenter: **Elise Bayraktar**, Universite Gustave Eiffel, France
Co-authors: Emmanuelle Clement

A stable Cox-Ingersoll-Ross (α-stable CIR) process defined by $dX_t = (a - bX_t)dt + \sigma X_t^{1/2} dW_t + \delta X_{t-}^{1/\alpha} dL_t^{\alpha}, \quad X_0 = x_0 > 0$, where (L_t^{α}) is a stable Levy process with non-negative jumps and jump activity index $\alpha \in (1,2)$ and (W_t) is a standard Brownian motion. The pure jump case $\sigma = 0$ is considered. The aim is to study the joint estimation of drift, scaling and jump activity parameters $(a, b\delta, \alpha)$ from high-frequency observations of the process on a fixed time period. The existence of a joint estimator of (a, b, δ, α) is proven based on an approximation of the likelihood function, which is consistent and asymptotically conditionally Gaussian. Moreover, the uniqueness of the drift estimators is established, assuming that δ and α are known or consistently estimated. Easy-to-implement preliminary estimators of all parameters are proposed, and those are improved by a one-step procedure. The conclusion is by proposing an estimation method for the general case $\sigma > 0$ based on the method of moments.

C0870: **Invariant density estimation of self-exciting jump-diffusion**
Presenter: **Chiara Amorino**, University Pompeu Fabra, Spain
Co-authors: Arnaud Gloter, Charlotte Dion, Sarah Sarah Lemler

Results are presented on estimating the invariant density associated with $(X_t, \lambda_t)_{t \geq 0}$, where X is diffusion with jumps driven by a multivariate nonlinear Hawkes process, and λ is a piecewise deterministic Markov process (PDMP) defining the stochastic intensity. Estimating the invariant density is crucial due to its applications in physics and numerical methods, particularly Markov chain Monte Carlo. Non-parametric estimation for the stationary measure of a continuous mixing process is a well-established yet evolving topic. Estimating the invariant density $\pi(x,y)$ of (X, λ) is proposed using kernel density estimation, assuming a continuous record of X is available. Accuracy is measured by the pointwise L^2 error, requiring pre-estimation of λ's parameters, yielding an estimator $\hat{\lambda}$, whose analysis is crucial for obtaining the main results. The main contributions include explicitly determining the convergence rates of the proposed estimator, which vary based on the estimation point. These results are compared to those for estimating the invariant density of a Levy process.

C1482: **Efficient estimation of stable-Levy SDEs with constant scale coefficient**
Presenter: **Thi Bao Tram Ngo**, University of Evry Val d Essonne, France
Co-authors: Alexandre Brouste, Laurent Denis

The joint parametric estimation of the drift coefficient, the scale coefficient, and the jump activity index in stochastic differential equations driven by a symmetric stable Levy process are considered based on high-frequency observations. Firstly, the LAMN property for the corresponding Euler-type scheme is proven, and lower bounds for the estimation risk in this setting are deduced. Therefore, when the approximation scheme experiment is asymptotically equivalent to the high-frequency observation of the solution of the considered stochastic differential equation, these bounds can be transferred. Secondly, since the maximum likelihood estimator can be time-consuming for large samples, an alternative to Le Cam's one-step procedure is proposed in the general setting. It is based on an initial guess estimator, which is a combination of generalized variations of the trajectory for the scale and the jump activity index parameters and a maximum likelihood type estimator for the drift parameter. This proposed one-step procedure is shown to be fast, asymptotically normal, and even asymptotically efficient when the scale coefficient is constant. In addition, the performances in terms of asymptotic variance and computation time on samples of finite size are illustrated with simulations.

C1557: **Bats monitoring: A classification procedure of bats behaviors based on Hawkes processes**
Presenter: **Romain Lacoste**, CNRS UMR 8050 Gustave Eiffel University, France
Co-authors: Christophe Denis, Charlotte Dion, Laure Sansonnet, Yves Bas

We are interested in classifying the commuting and foraging behaviour of bats at delimited geographical areas, namely sites, throughout France. To predict the majority behaviour on these sites, echolocation call data is used, recorded as part of Vigie-Chiro participatory project. As the temporal distribution of calls is a relevant indicator of behaviour, providing an adequate model of this distribution is a matter of great interest. Given the self-exciting dynamics observed in foraging behaviour, we propose to model bat calls by Hawkes processes. Specifically, the start time of each call emitted on a site is considered an event of a Hawkes process. Taking advantage of this modelling, a suitable procedure is used that relies on the empirical risk minimization principle to discriminate between the two classes. Then, the performance of the procedure is assessed on synthetic data through comprehensive numerical experiments. The overall methodology is evaluated with a goodness-of-fit test. Finally, the results obtained are presented in the real data set. The classification results are convincing, and the relevance of the method is shown, which could contribute to a better understanding of behavioral determinants and open up broad perspectives in spatial ecology.

CO171 Room S-1.06 INFERENCE OF COMPLEX STOCHASTIC DYNAMIC MODELS Chair: Danna Zhang

C0643: **Online inference for stochastic gradient descent with dropout regularization**
Presenter: **Jiaqi Li**, University of Chicago, United States
Co-authors: Johannes Schmidt-Hieber, Wei Biao Wu

An online inference method is proposed for the stochastic gradient descent (SGD) iterates with dropout regularization in linear regression. Specifically, the geometric-moment contraction (GMC) is established for constant step-size SGD dropout iterates to show the existence of a unique stationary solution to the dropout recursive function. By the GMC property, quenched central limit theorems (CLT) are provided for the difference between dropout and ℓ^2-regularized iterates, regardless of the fixed initial points. The CLT for the difference between the Ruppert-Polyak averaged SGD (ASGD) with dropout and ℓ^2-regularized iterates is also presented. Based on these asymptotic normality results, an online estimator is further introduced for the long-run covariance matrix of ASGD dropout to facilitate inference recursively with efficiency in computational time and memory. Numerical experiments also demonstrated that the proposed confidence intervals for ASGD dropouts can achieve the desired asymptotic coverage probability.

C0910: **Uncertainty quantification with a latent variable model**
Presenter: **Mengyu Xu**, University of Central Florida, United States

To understand the behavior of a complex system, one is often interested in some key internal quantities that are not directly observable. The aim is to study the inference of the final output and the internal quantities of a complex system from controlled system inputs and multi-fidelity realizations of the latent internal quantities. A latent variable model is studied under a Bayesian framework. The model is two-step: (1) the inference of the latent internal variable given their noisy approximations and the system inputs; and (2) the study of the system outputs from the inferred internal variables. Linear and nonlinear approximations are employed in the second step. For the nonlinear approximation, Markov Chain Monte Carlo is employed. In addition, the inverse problem is studied, i.e., estimate the posterior of the internal quantities from their noisy measurements and the system inputs and output. This provides insight into the system's application of fault detection. The approach is verified against a numerical model, demonstrating its veracity.

C0951: **Tensor-augmented transformers for multi-dimensional time series forecasting**
Presenter: **Yuefeng Han**, University of Notre Dame, United States

Multi-dimensional time series data, such as matrix and tensor-valued time series, are increasingly prevalent in fields such as economics, finance, and climate science. Traditional transformer models, though adept with sequential data, do not effectively preserve these multi-dimensional structures, as their internal operations, in effect, flatten multi-dimensional observations into vectors, thereby losing critical multi-dimensional relationships and patterns. To address this, the tensor-augmented transformer (TEAFormer) is introduced, a novel method that incorporates tensor expansion and compression within the transformer framework to maintain and leverage the inherent multi-dimensional structures, thus reducing computational costs and improving prediction accuracy. The comprehensive experiments, which integrate the TEA module into three popular time series transformer models across three real-world benchmarks, show significant performance enhancements, highlighting the potential of TEAFormers for cutting-edge time series forecasting.

C1004: **Linear discriminant analysis of high-dimensional time series**
Presenter: **Danna Zhang**, University of California, San Diego, United States

Classification is one of the fundamental problems in time series analysis, where the goal is to assign a new observed series to one of multiple known classes. While the classification of low-dimensional time series has been well studied, the investigation of high-dimensional cases remains limited. Sparse linear discriminant analysis is applied to high-dimensional time series, and conditions for the consistency of the time series LDA rule are established for both Gaussian and non-Gaussian processes. Numerical studies and application to fMRI data are conducted to corroborate the results.

CO273 Room S-1.27 ADVANCES IN ANALYZING HIGH DIMENSIONAL DATA Chair: Hossein Moradi Rekabdar.

C0635: **A multi-bin rarefying method for evaluating alpha diversities in TCR sequencing data**
Presenter: **Mo Li**, UL at Lafayette, United States

The purpose is to examine the impact of library size variation on T cell receptor (TCR) diversity analysis, specifically evaluating the commonly used overall rarefying method. Library sizes, reflected in sequencing experiment reads, vary significantly across samples, complicating alpha diversity estimation and comparison. Despite its common use, the overall rarefying method's effectiveness remains unverified. An innovative multi-bin rarefying approach is developed that partitions samples into multiple bins according to their library sizes, conducts rarefying within each bin for alpha diversity calculations, and performs meta-analysis across bins. Extensive simulations using real-world cytomegalovirus (CMV) data highlight the inadequacy of the overall rarefying approach in controlling the confounding effect of library size. The method proves robust in addressing library size confounding, outperforming competing normalization strategies by achieving better-controlled type-I error rates and enhanced statistical power in association tests.

C0703: **Trigonometry-transformation based correlation coefficient with an application to sufficient variable selection**
Presenter: **Pei Wang**, Bowling Green State University, United States

The technique of variable selection has gained widespread popularity for reducing data size, particularly in the context of large p small n datasets. A novel criterion is introduced based on the correlation coefficient derived from trigonometry transformations. This innovative criterion serves as a metric for assessing the relationship between the response and each predictor. When integrated into a two-step selection procedure, it becomes a valuable tool for variable selection. Notably, this approach is model-free, providing robustness against model misspecification. The asymptotic and sure selection properties are established, and the effectiveness of the proposed method is demonstrated through extensive numerical studies and real data analysis.

C0997: **Enhancing variable selection with preliminary sufficient dimension reduction in semi-competing risks data analysis**
Presenter: **Chenlu Ke**, Virginia Commonwealth University, United States

A new framework is introduced with an efficient algorithm for feature screening in the challenging context of ultrahigh dimensional semi-competing risk data. Specifically, the two-stage procedure initially employs a dual screening mechanism to select a coarse set of features that are potentially relevant to both terminal and nonterminal endpoints. This leads to the estimation of the augmented central subspace, pivotal for both endpoints and censoring, based on the selected features. In the second stage, refined sets of important features for the nonterminal and terminal events are further identified using an inverse probability-of-censoring weighted filter, where the central subspace estimator is used to obtain the weights adjusting for censoring. The proposed framework is model-free, and it does not require independent censoring. Asymptotic properties are established under minor assumptions. The promising performance of the proposed method is demonstrated through simulations and gene expression data analysis.

C1022: **On partial envelope approach for modeling spatial-temporally dependent data**
Presenter: **Wenbo Wu**, University of Texas at San Antonio, United States

In the new era of big data, modeling multivariate spatial-temporally dependent data is a challenging task due to not only the high dimensionality of the features but also complex spatial-temporal associations among the observations across different locations and time points. To improve the estimation efficiency, a spatial-temporal partial envelope model is proposed, which is parsimonious and effective in modeling high-dimensional spatial-temporal data. The partial envelope model is proposed under a linear coregionalization model framework, which allows for a heterogeneous spatial-temporal covariance structure for different components of the response vector. The maximum likelihood estimator for the proposed model can be obtained through a Grassmann manifold optimization. An asymptotic result is obtained for the estimator, and thorough simulation studies are conducted to demonstrate the soundness and effectiveness of the proposed method. The proposed model is also applied to analyze the crowd-sourcing weather data collected from personal weather stations in the city of Syracuse, NY, United States.

CO060 Room K0.16 ADVANCES IN LONGITUDINAL STUDIES Chair: Sanjoy Sinha

C0340: **Estimation of dynamic Logit mixed models for multinomial responses with categorical covariates**
Presenter: **Alwell Oyet**, Memorial University, Canada
Co-authors: Brajendra Sutradhar

A situation is considered where multinomial responses are collected repeatedly over a short period of time from a large number of independent individuals along with individuals' categorical covariate information. Dynamic logit models are developed under the assumption that the responses are influenced by (a) The categorical covariates, (b) Past multinomial responses, and (c) Some category-prone unobservable variables. By category prone, unobservable variables are referred to that influence the responses into specific categories. The likelihood estimation of the effects of the covariates, the dynamic dependence parameters, and the variances of the category-prone random effects are discussed. The results of simulation studies in special cases of the longitudinal mixed model, namely, a cross-sectional model, longitudinal fixed effects model and a longitudinal mixed model, are discussed. Asymptotic properties of the covariate effects parameters may be discussed if time permits.

C0748: **Leveraging longitudinal data for enhanced survival analysis using a novel deep transformer model**
Presenter: **Pingzhao Hu**, Western University, Canada

Traditional statistical methods, while capable of analyzing the quality of life and toxicity data individually, often struggle with the efficiency and integration required for comprehensive multi-dataset analyses in clinical settings. This limitation hampers the long-term understanding of patients. This research develops an innovative model using a modified transformer encoder framework enhanced with an attention-free transformer (AFT). This model is engineered to concurrently process and correlate clinical variables with longitudinal assessments of quality of life and toxicity. Utilizing a transformer encoder architecture optimized with AFT, time-series clinical data is efficiently processed. This framework is integrated with Cox regression analysis, and measured via the concordance index. A key highlight of the model is its adeptness at integrating diverse datasets, including cross-sectional and longitudinal datasets, while adeptly managing variations in measurement times and frequencies across different subjects. Tested on a longitudinal dataset of 750 colorectal cancer patients over a two-year period, the model outstripped conventional analytical methods in predictive accuracy and integration, as evidenced by a substantial enhancement in the concordance index. The findings highlight the models potential to transform clinical decision-making processes, setting the stage for further exploration into its clinical implications and adaptability.

C1619: **Nonparametric treatment model smoothing under sparsity for causal inference with longitudinal treatments**
Presenter: **Mireille Schnitzer**, Universite de Montreal, Canada
Co-authors: David Berger, Yan Liu, David Benkeser

Marginal structural models (MSM) for longitudinal treatments are often fit using estimates of the probability of treatment at each time point, conditional on covariate history. Such approaches are limited by data sparsity in the treatment model, i.e. practical positivity violations, which can greatly increase estimation variance. However, it is possible to smooth model features, including removing covariates that have no relationship with the outcome, in order to retain adjustment for confounding while sometimes greatly decreasing variance. A longitudinal outcome adaptive LASSO is previously proposed to perform covariate reduction in working parametric treatment models, which was successful in theory and practice at reducing estimation variance under structural and modeling assumptions. However, in addition to relying on parametric models, this approach does not allow for uniform convergence of the estimation of the treatment model functions, precluding inference for the MSM parameters. Therefore, the proposal is to extend the nonparametric outcome-highly-adaptive LASSO to longitudinal treatments, obtaining a regular asymptotically linear estimator under data-adaptive learning of the outcome model, under purposeful misspecification of the treatment model.

C0333: **Joint analysis of longitudinal count and binary data with outliers**
Presenter: **Sanjoy Sinha**, Carleton University, Canada

The purpose is to discuss an innovative, robust method for jointly analyzing longitudinal count and binary responses. The method is useful for bounding the influence of potential outliers when estimating the model parameters. The asymptotic properties of the proposed robust estimators will be discussed. The empirical properties are investigated based on a simulation study. The proposed method is illustrated using some real data from a public health study.

CO288 Room K0.18 ADVANCES IN MINIMAX OPTIMALITY Chair: Cecile Durot

C0481: **Minimax estimation in the functional regression model with a functional output**
Presenter: **Gaelle Chagny**, CNRS, Universite de Rouen Normandie, France
Co-authors: Anouar Meynaoui, Angelina Roche

The problem of nonparametric estimation of a linear regression model is addressed, where both the covariate and the response variable are functional random variables. Projection estimators for the conditional expectation operator are introduced. Their prediction risk achieves a non-asymptotic sharp upper-bound as a classical bias-variance compromise. Then, the automatic trade-off is realized thanks to a model selection device (penalized criterion). An oracle-type inequality is proved, and convergence rates are derived from ellipsoidal regularity spaces. They match with a lower-bound (also proved), and thus, the procedure is optimal in the adaptive and in the minimax sense. A numerical study (over simulated data and over a real-data set) is also presented.

C0485: **On optimal adaptive estimation of a density**
Presenter: **Mathieu Sart**, Universite Jean Monnet, France

The problem of estimating a density f is tackled on the real line $\mathbb{R}$. A new way of thresholding the coefficients in wavelet methods is presented. The risk of the estimator is evaluated using a global $\mathbb{L}^1$ risk and relying on the minimax approach. The assumptions made about the density are mild in the sense that f may have some spatial variability, may not be bounded, continuous, compactly supported or in $\mathbb{L}^2$. The advantage of the method discussed is that it avoids the undesirable log factors that usually appear in older procedures. Particular attention is paid to assumptions about the tails of the distribution and their impact on the optimal estimation rates. New ones are thus revealed.

C0567: **Minimax geometric inference: Open and closed problems**
Presenter: **Catherine Aaron**, Universite Clermont Auvergne, France

An overview of minimax set and manifold estimators are presented from convex sets to stratified manifold estimators. The convergence of the Hausdorff distance between the set and its estimator is first discussed. Another way is proposed to measure the error in set estimation, such as Wasserstein distance or integral probability metric. The more challenging problem of minimax volume estimation and the still open problem of minimax perimeter estimation are also discussed.

C1043: **Minimax optimal rates of convergence in the shuffled and unlinked regression, and deconvolution under vanishing noise**
Presenter: **Debarghya Mukherjee**, Princeton University, United States
Co-authors: Cecile Durot

Shuffled and unlinked regression represent intriguing challenges that have garnered considerable attention in many fields, including but not limited to ecological regression, multi-target tracking problems, image denoising, etc. However, a notable gap exists in the existing literature, particularly in vanishing noise, i.e., how the estimation rate of the underlying signal scales with the error variance. The aim is to bridge this gap by delving into the monotone function estimation problem under vanishing noise variance, i.e., the error variance is allowed to go to 0 as the number of observations increases. The investigation reveals that, asymptotically, the shuffled regression problem exhibits a comparatively simpler nature than the unlinked regression; if the error variance is smaller than a threshold, then the minimax risk of the shuffled regression is smaller than that of the unlinked regression. On the other hand, the minimax estimation error is of the same order in the two problems if the noise level is larger than that threshold. The analysis is quite general; any smoothness of the underlying monotone link function is not assumed. Because these problems are related to deconvolution, bounds for deconvolution in a similar context are also provided. Through this exploration, the contribution is to understand the intricate relationships between these statistical problems and shed light on their behaviors when subjected to the nuanced constraint of vanishing noise.

CO207 Room K0.19 LATEST TRENDS IN CLUSTERING AND CLASSIFICATION OF COMPLEX DATA I Chair: Marta Nai Ruscone

C1426: **Kernel metric learning for mixed-type fuzzy clustering**
Presenter: **John Thompson**, University of British Columbia, Canada
Co-authors: Daniel Krasnov

Fuzzy clustering algorithms, such as the popular fuzzy C-means algorithm, extend hard clustering to allow for a degree of uncertainty in the cluster assignments through a fuzzifying parameter in the objective function. However, challenges remain in selecting an optimal fuzzy parameter, incorporating mixed continuous and categorical data types, and balancing variables based on their importance to fuzzy clustering. A kernel weighting approach is proposed in the objective function, where bandwidth selection controls variable importance and reduces the influence of selecting a fuzzy parameter. Methods are discussed for selecting bandwidths through kernel metric learning, as well as bandwidth optimization in the objective function. The kernel metric is incorporated into the fuzzy C-means objective function, and it is shown that the bandwidth and the fuzzifying parameter are correlated. It is found that the overall effect of fuzzifying parameter choice on cluster center values and fuzzy adjusted Rand index is mitigated by selecting bandwidths that optimize the objective function. This method is applied to simulated and real mixed-type benchmark datasets to demonstrate improvements in clustering performance against current fuzzy clustering algorithms.

C1439: **Statistical few-shot learning via parameter pooling**
Presenter: **Semhar Michael**, South Dakota State University, United States
Co-authors: Andrew Simpson

The emergence of high-dimensional data characterized by the observations being partitioned into many classes with a limited number of samples per class brings a significant challenge to classical probabilistic machine-learning techniques. Few- or one-shot learning problems are among these classes of problems. Given the nature of the few-shot learning framework, strong assumptions about the data-generating process must be made. To get non-singular and stable estimates for the covariance matrices of each class, it is often assumed that each class has the same covariance matrix as in linear discriminant analysis (LDA). In this framework, given that the number of classes tends towards infinity in this setting, this assumption is extreme. The strong assumption of LDA is relaxed, and stable estimates of the covariance matrices are obtained. In this regard, a finite number of distributions is assumed to exist from which a class can come. In the case of Gaussian distributions, this amounts to assuming a finite number of covariance matrices a class can take, which is less than the number of classes. Using simulation studies and real data analysis will demonstrate the utility of the proposed methodology.

C1465: **On the use of contaminated Gaussian distributions for modeling heavy tails and outliers**
Presenter: **Yana Melnykov**, The University of Alabama, United States

Gaussian mixture modeling is popular among researchers and practitioners due to its interpretability and relatively straightforward mathematical handling. However, using Gaussian mixtures can be problematic when dealing with outliers and heavy-tailed data groups. The existing literature offers several methods to tackle these issues, with one prominent approach involving the use of contaminated normal distributions. These distributions represent a mixture of two normal components with a common location parameter and one scale parameter being the multiple of the other one. This model allows improved capturing of the potentially heavy distribution tails. Another popular use of contaminated normal distributions is to detect mild outliers. The analysis of both applications of contaminated normal distributions is considered, providing novel insights into the use of this useful model.

C1642: **Rank-based strategies for clustering distributions of pairwise scores**
Presenter: **Christopher Saunders**, South Dakota State Univerisity, United States
Co-authors: Janean Hanka, Clarissa Giefer

A frequently encountered problem in few-shot learning and forensic source identification is to assign a query object to a class of objects with respect to a score (or sometimes a metric) function based only on a set of pairwise similarities between the objects. In these problems, a metric distinguishes between same-and different-class comparisons. This metric is then used to construct a one nearest neighbor classifier. Unfortunately, a more sophisticated class of models or classifiers is impractical because a few number of observations per class limits the ability to properly estimate the induced joint distribution of a set of scores. A potential solution for this limitation would be to cluster classes of objects if their within-class distributions with respect to the learned metric are sufficiently similar. Strategies for pooling within-class comparisons that have the same within-class distributions are developed. When considering the set of within-class comparisons, this is a U-process with respect to the kernel induced by the learned metric. Goodness-of-fit and rank-based level-alpha tests are proposed for measuring the degree of dissimilarity of these two sets of distributions of scores that will account for the U-process dependency. Finally, strategies for combining the pairs of tests between the classes of objects are developed using local false discovery mixture models to make statements concerning which sets of classes share the same within-class distribution.

CO130 Room K0.20 RECENT ADVANCES IN QUANTILE AND M-QUANTILE REGRESSION Chair: Luca Merlo

C0626: **A Hausman-type test for detecting constancy of quantile slopes**
Presenter: **Jayeeta Bhattacharya**, University of Southampton, United Kingdom

A test of constancy of quantile regression (QR) slope parameters is proposed across a range of quantile levels. The proposed test is inspired by a past study's testing idea, and two QR estimators are compared, the unconstrained and constrained one (obtained by imposing the constancy restriction), which should converge to the same limit under the null hypothesis of correct constancy specification, and diverge under the alternative. The critical values for the test are obtained using random weighting bootstrap. The size and power of the test infinite samples are analyzed through Monte Carlo experiments.

C0796: **Extreme conditional quantile estimation for location-scale regression models and time series**
Presenter: **Gilles Stupfler**, University of Angers, France
Co-authors: Marco Oesting

Motivated by applied questions in environmental science, finance and insurance, such as the prediction of the magnitude of a potential extreme rainfall event tomorrow given weather parameters today, or the prediction of the value of large losses on a financial asset given the overall state of a financial market, we construct conditional extreme quantile estimators in location-scale regression models based on residuals from a preliminary estimation of model structure. The crucial difficulty in order to work out the asymptotic behavior of the resulting estimators is that residuals will typically not be independent or even identically distributed, even in simple models such as linear regression. Recent work has shown that residual-based versions of extreme value estimators are consistent and asymptotically normal, just as their unachievable true error-based counterparts would be, provided the residuals are in some sense uniformly close to the corresponding regression errors. It is shown that this assumption can be substantially weakened by taking a different route, not relying on the validity of a Gaussian approximation to the so-called tail empirical process, thus leading to a theoretical framework that can, in particular, handle a large array of classical time series models without having to impose unnecessary technical restrictions.

C0986: **M-quantile regression shrinkage and selection via the Lasso and Elastic Net**
Presenter: **Francesco Pantalone**, University of Southampton, United Kingdom
Co-authors: Maria Giovanna Ranalli, Nicola Salvati, Lea Petrella

An M-quantile regression model with Lasso and Elastic Net penalizations is presented. This new methodology allows (i) to identify the best predictors via model selection, (ii) to investigate the relationship between response and covariates at different M-quantiles of the conditional response distribution, and (iii) to be robust to the presence of outliers. Finally, heterogeneity in the data can be accounted for via B-spline. A real application of the effect of traffic on air quality in the city of Perugia (Italy) is presented.

C0907: **Hidden Markov linear quantile graphical model**
Presenter: **Beatrice Foroni**, University of Pisa, Dip. Economia e Management, Italy
Co-authors: Luca Merlo, Nicola Salvati, Lea Petrella

Graphical models are crucial for understanding interdependencies among multiple variables in fields such as genome biology, finance, and environmental studies. These data often evolve over time and are influenced by hidden variables, necessitating models that capture these temporal dynamics. Hidden Markov models (HMMs) are particularly suited for this purpose. Previous work has explored the estimation of graphical models within HMMs using multivariate Gaussian emission distributions. However, the assumption of Gaussianity is often unrealistic for many practical applications. To address the need for modeling time-varying conditional dependencies without the Gaussian assumption, a sparse hidden Markov linear quantile graphical model (HMLQGM) is proposed. This approach leverages the conditional quantile to infer conditional independence structures. Parameter estimation is achieved through an expectation-maximization (EM) algorithm combined with a LASSO penalty, which facilitates the identification of the most relevant connections in the graph structure. Simulation studies demonstrate that HMLQGM effectively recovers dependency structures across various scenarios, highlighting its potential for broad applicability in analyzing complex, temporally evolving data.

CO249 Room K0.50 SUBSAMPLING AND NON-LINEAR PROBLEMS: NEW PROPOSALS IN OPTIMAL DESIGN Chair: Chiara Tommasi

C0969: **Subdata selection for prediction under model misspecification**
Presenter: **Alvaro Cia-Mina**, University of Navarra, Spain
Co-authors: Laura Deldossi, Jesus Lopez-Fidalgo, Chiara Tommasi

Subsampling is commonly employed to improve computational efficiency in regression models. However, existing methods primarily focus on minimizing errors in estimating parameters, whereas the main practical goal of statistical models often lies in minimizing prediction errors. A novel approach to selecting subdata for linear models under model misspecification is introduced. The method considers the distribution of covariates and specifically addresses scenarios with large samples where obtaining labels for the response variable is costly. A strategy is proposed to minimize the bias term of the random-X prediction error. As anticipated based on theoretical considerations, the method demonstrates a reduction in the bias of the prediction mean squared error compared to existing methods. Through simulations, empirical evidence of the performance and potential of the approach in enhancing prediction accuracy under model misspecification is presented.

C0660: **Optimal design for parameters estimation in generalized linear models with treatment-by-covariate interactions**
Presenter: **Rosamarie Frieri**, University of Bologna, Italy
Co-authors: Alessandro Baldi Antognini, Maroussa Zagoraiou

In an experiment aimed at comparing multiple treatments, especially in the clinical context, subject covariate information is often available and should be incorporated not only in the data analysis at the end of the study but also into the treatment allocation scheme. With the advances in medical research and the advent of precision medicine, scientists have identified many new covariates/biomarkers that may be linked with certain diseases and could strongly influence patients' responses to treatments. The D- and A-optimal designs are derived for parameter estimation for generalized linear models with treatment-by-covariate interactions. The optimal designs require a set of equality constraints involving i) the unknown model parameters and ii) the subject covariate values. Since such conditions are not directly attainable, the optimum can be achieved sequentially by adopting a novel class of covariate-adjusted response-adaptive randomization, which aims at minimizing, at each step of the sequential procedure, the Euclidean distance between the current allocation and the optimum. The performance of the proposed approach in terms of estimation efficiency is assessed both theoretically and through an extensive simulation study.

C0867: **Subsampling and its advantages for exponential family models**
Presenter: **Subhadra Dasgupta**, Ruhr University Bochum, Germany
Co-authors: Holger Dette

A novel two-stage subsampling algorithm is proposed based on optimal design principles. In the first stage, a density-based clustering algorithm and a Markov chain Monte Carlo method are used to identify an approximating design space for the predictors from an initial subsample. Next, an optimal approximate design is determined on this design space. Finally, matrix distances, such as the Procrustes, Frobenius, and square-root distance, are used to define the remaining subsample so that its points are closest to the support points of the optimal design. The approach reflects the specific nature of the information matrix as a weighted sum of non-negative definite Fisher information matrices evaluated at the design points and applies to a large class of regression models, including models where the Fisher information is of rank larger than 1. Additionally, the algorithm removes outliers from the subsample, leading to reliable predictions.

C0394: **Optimal discrimination designs for a constrained type of random effects models**
Presenter: **Sergio Pozuelo Campos**, University of Castilla-La Mancha, Spain
Co-authors: Victor Casero-Alonso, Jesus Lopez-Fidalgo, Chiara Tommasi, WengKee Wong

It is common in the optimal design literature to assume that the model has no random effects and it is known, apart from model parameters. Random effects models are widely applied across all disciplines, particularly in the life sciences and clinical studies. The assumption is that there are several plausible models, and the aim is to find a design that optimally discriminates among models with random effects. A common design criterion for discriminating between models with fixed effects is the Kullback-Leibler divergence criterion. It is a maximin-type of criterion, not differentiable and is a 2 or 3-layer nested optimization problem over very distinct domains. Consequently, optimal designs for discriminating models are notoriously difficult to determine, not only analytically but also computationally. Theoretical results are provided for the KL-optimality criterion value for discriminating among random effects models, and a nature-inspired metaheuristic algorithm is implemented to facilitate the search for an optimal discrimination design. The methodology is quite general and applies to discriminating random effects models with multiple interacting experimental conditions, which may be continuous or discrete. Two applications are provided; the first finds a design that optimally discriminates among fractional polynomials with a single continuous variable, and the second identifies the best design to discriminate among several multi-factor random effects models.

CO285 Room K2.31 (Nash Lec. Theatre) EXPLORING NEW FRONTIERS IN CAUSAL MEDIATION ANALYSIS Chair: Yi Zhao

C0638: **Mediation analysis with high dimensional exposures and confounders**
Presenter: **Qi Zhang**, University of New Hampshire, United States

High-dimensional mediation analysis has been receiving increasing popularity, largely motivated by the scientific problems in genomics and biomedical imaging. Previous literature has primarily focused on mediator selection for high-dimensional mediators. The aim is to estimate and infer the overall indirect effect of high dimensional exposures and high dimensional mediators. MedDiC, a novel debiased estimator of the high dimensional overall indirect effect, is proposed based on the difference-in-coefficients approach. The proposed method is evaluated using intensive simulations, and it is found that MedDiC provides valid inference and offers higher power and shorter computing time than the competitors for both low-dimensional and high-dimensional exposures. MedDiC is also applied to a mouse f2 dataset for diabetes study and a dataset composed of diverse maize inbred lines for flowering time, and MedDiCyields is shown to provide more biologically meaningful gene lists, and the results are reproducible across analyses using different measures of identical biological signal or related phenotype as the outcome.

C0802: **Semiparametric causal mediation analysis in cluster-randomized experiments**
Presenter: **Fan Li**, Yale University, United States
Co-authors: Chao Cheng

In cluster-randomized experiments, there is emerging interest in exploring the causal mechanism in which a cluster-level treatment affects the outcome through an intermediate outcome. The formal semiparametric efficiency theory is developed to motivate several doubly robust methods for addressing several mediation effect estimands corresponding to both the cluster-average and the individual-level treatment effects in cluster-randomized experiments: the natural indirect effect, natural direct effect, and spillover mediation effect. The efficient influence function is derived for each mediation effect, and carefully parameterize each efficient influence function to motivate practical strategies for operationalizing each estimator. Parametric working models and data-adaptive machine learners are considered to estimate the nuisance functions and obtain semiparametric efficient causal mediation estimators in the latter case. The methods are illustrated via extensive simulations and two completed cluster-randomized experiments.

C1075: **Mediation analysis with graph mediator**
Presenter: **Yi Zhao**, Indiana University, United States
Co-authors: Yixi Xu, Yi Zhao

A mediation analysis framework, when the mediator is a graph, is introduced. A Gaussian covariance graph model is assumed for graph presentation. Causal estimands and assumptions are discussed. With a covariance matrix as the mediator, a low-rank representation is introduced, and parametric mediation models are considered under the structural equation modelling framework. Assuming Gaussian random errors, likelihood-based estimators are introduced to simultaneously identify the low-rank representation and causal parameters. An efficient computational algorithm is proposed, and asymptotic properties of the estimators are investigated. Via simulation studies, the performance of the proposed approach is evaluated. Applying a resting-state fMRI study, a brain network is identified within which functional connectivity mediates the sex difference in the performance of a motor task.

C1053: **Nonparametric mediation estimators that accommodate multiple mediators and multiple intermediate confounders**
Presenter: **Kara Rudolph**, Columbia University, United States

Mediation analysis is appealing for its ability to improve understanding of the mechanistic drivers of causal effects, but real-world data complexities challenge its successful implementation, including 1) the existence of post-exposure variables that also affect mediators and outcomes (thus confounding the mediator-outcome relationship), that may also be 2) multivariate, and 3) the existence of multivariate mediators. All three challenges are present in the mediation analysis considered. Interventional direct and indirect effects (IDE/IIE) accommodate post-exposure variables that confound the mediator-outcome relationship, but currently, no readily implementable nonparametric estimator for IDE/IIE exists that allows for both multivariate mediators and multivariate post-exposure intermediate confounders. This gap is addressed by extending a recently developed nonparametric estimator for the IDE/IIE to allow for multivariate mediators and multivariate post-exposure confounders simultaneously. The proposed estimation approach is applied to the analysis, including walking through a strategy to account for other, possibly co-occurring intermediate variables when considering each mediator subgroup separately.

CO119 Room K2.40 FRONTIERS IN STATISTICAL NETWORK ANALYSIS Chair: Keith Levin

C0165: **Detection and statistical inference on informative core and periphery structures in weighted directed networks**
Presenter: **Wen Zhou**, New York University, United States
Co-authors: Tianxi Li, Wenqin Du

In network analysis, noises and biases due to peripheral or non-essential components can mask pivotal structures and hinder the efficacy of many network modelling and inference procedures. Recognizing this, the identification of the core-periphery (CP) structure has emerged as a crucial data pre-processing step, while the efforts to detect CP for directed weighted networks have been limited. Existing efforts either fail to account for the directionality or lack the theoretical justification of the identification procedure. The aim is to answer three pressing questions: (i) How can informative and non-informative structures in weighted directed networks be distinguished? (ii) What approach offers computational efficiency in discerning these components? (iii) Upon the detection of CP structure, can uncertainty be quantified to evaluate the detection? The signal-plus-noise model is adopted, categorizing uniform relational patterns as non-informative, by which we define the sender and receiver peripheries. Furthermore, instead of confining the core component to a specific structure, it is considered complementary to either the sender or receiver peripheries. Based on the definitions of the sender and receiver peripheries, spectral algorithms are proposed to identify the CP structure in directed weighted networks. The algorithm stands out with statistical guarantees, ensuring the identification of sender and receiver peripheries with overwhelming probability.

C0271: **Modelling sparse influence networks with Hawkes process while controlling for global influence**
Presenter: **Alexander Kreiss**, Leipzig University, Germany
Co-authors: Enno Mammen, Wolfgang Polonik

Vertices are considered in a network who are able to cast events. The neighbors of a given vertex are supposed to take note of the events cast by this vertex and change their own behavior accordingly. The model believes, more precisely, that the activity of one vertex increases the activity of its neighbors. This is called peer effects. However, there might also be other (observed) information which increases or decreases the activity of the vertices. This is called global effects. A Hawkes model is seen, which incorporates both peer and global effects. This allows for the estimation of the network, that is, the influence structure while controlling for network effects or the estimation of the global effects while controlling for peer effects. The estimation is based on a LASSO strategy, which respects sparsity in the network.

C0510: **Estimation of the number of communities for sparse networks**
Presenter: **Shirshendu Chatterjee**, City University of New York, United States
Co-authors: Sharmodeep Bhattacharyya, Neil Hwang, Jiarui Xu

Among the nonparametric methods of estimating the number of communities (K) in a community detection problem, methods based on the spectrum of the Bethe Hessian matrices have garnered much popularity for their simplicity, computational efficiency, and robustness to the sparsity of data. For certain heuristic choices of these matrices, such methods were shown to be consistent for semi-dense networks. A consistent K-estimation procedure is developed based on spectral properties of suitable Bethe Hessian matrices in the sparse regime. The performance of the resulting estimation procedure is evaluated theoretically and empirically through extensive simulation studies and application to a comprehensive collection of real-world network data.

C1026: **Optimizing the induced correlation in omnibus joint graph embeddings**
Presenter: **Vince Lyzinski**, University of Maryland, College Park, United States
Co-authors: Konstantinos Pantazis, Michael Trosset, William Frost, Carey Priebe

Theoretical and empirical evidence suggests that joint graph embedding algorithms induce correlation across the networks in the embedding space. In the Omnibus joint graph embedding framework, previous results explicitly delineated the dual effects of the algorithm-induced and model-inherent correlations on the correlation across the embedded networks. This algorithm-induced correlation is key to subsequent inference, as sub-optimal Omnibus matrix constructions can lead to a loss in inference fidelity. We present the first efforts to automate the Omnibus construction in order to address two key questions in this joint embedding framework: the correlation–to–OMNI problem and the flat correlation problem. In the flat correlation problem, we seek to understand the minimum algorithm-induced flat correlation (i.e., the same across all graph pairs) produced by a generalized Omnibus embedding. Working in a subspace of the fully general Omnibus matrices, we prove a lower bound for this flat correlation and that the classical Omnibus construction induces the maximal flat correlation. In the correlation–to–OMNI problem, we present an algorithm named corr2Omni that, from estimated pairwise graph correlations, estimates the generalized Omnibus weights that induce optimal correlation in the embedding space. Moreover, in both simulated and real data settings, we demonstrate the increased effectiveness of our corr2Omni algorithm versus the classical Omnibus construction.

CO319 Room K2.41 SURVIVAL ANALYSIS: CENSORING AND COMPETING RISKS Chair: Takeshi Emura

C0170: A copula duration model with dependent states and spells

Presenter: **Ralf Wilke**, Copenhagen Business School, Denmark
Co-authors: Ming Sum Simon Lo, Shuolin Shi

A nested Archimedean copula model for dependent states and spells is introduced, and the link to a classical survival model with frailties is established. The model relaxes an important restriction of classical survival models as unobservable heterogeneities are permitted to be correlated with the observable covariates. Its modular structure has practical advantages as the different components can be separately specified, and estimation can be done sequentially or separately. This makes the model versatile and adaptable in empirical work. An application to labor market transitions with linked administrative data supports the need for a flexible specification of the dependence structure and the model for marginal survivals. The conventional Markov chain model is shown to give sizeable biased results in the application.

C0173: Survival analysis for matched health databases

Presenter: **Valerie Gares**, Univ Rennes, INSA, CNRS, IRMAR - UMR 6625, Rennes, France
Co-authors: Jean-Francois Dupuy, Vanessa Chezeu

The purpose is to investigate estimation in the Cox proportional hazards model from matched health databases. The situation where the explanatory variables and individual lifetimes are not reported in the same database is considered. A prior process of probabilistic record linkage is therefore necessary to obtain a complete database. A partial likelihood equation is proposed that incorporates unobserved variables whose distribution is determined by the record linkage process. Parameter estimation was conducted using the EM algorithm. The asymptotic properties of the estimators that are proposed are studied, and their performance is validated through simulations.

C0859: Effect measures for comparing consecutive survival times

Presenter: **Dennis Dobler**, TU Dortmund University, Germany
Co-authors: Marc Ditzhaus, Merle Munko, Dominic Edelmann, Simon Mack

The progression-free-survival ratio (PFSr) is a popular endpoint in oncology trials, where patients serve as their own control. Various inference methods have been proposed for this endpoint. However, many of them rely on unrealistic assumptions or fail to incorporate censoring correctly, leading to biased results. The close connection between the PFSr and the relative treatment effect is pointed out. Recent results about the latter are applied to multivariate survival data based on techniques from competing risk analysis. In this way, a mathematically correct inference strategy is developed for the PFSr under random right-censoring. To further consider the rising importance of interpretable estimands in medical research, effect measures are proposed based on the restricted mean survival times as an alternative in this context. For both endpoints, valid resampling procedures are presented to improve the performance for small sample sizes. The novel methods are exemplified by an extensive simulation study and a real data analysis.

C1508: Variable selection in high-dimensional survival analysis

Presenter: **Pilar Gonzalez-Barquero**, Universidad Carlos III de Madrid, Spain
Co-authors: Rosa Lillo, Alvaro Mendez-Civieta

The rise of high-dimensional datasets, characterized by a high number of covariates, introduces significant challenges to traditional models. These large datasets, while rich in information, complicate the decision-making process. In this context, variable selection methods are necessary to reduce dimensionality and make the problem feasible. The focus is on survival analysis, particularly on the performance of Cox proportional hazards models in high-dimensional settings with a significant proportion of censored data. In such scenarios, the model presents an infinite number of possible solutions for the regression coefficients, requiring regularization techniques like Lasso and adaptive Lasso. Various methods are proposed and evaluated for determining adaptive Lasso weights, including principal component analysis, ridge regression, univariate Cox regression, and the random survival forest algorithm. Additionally, these methods are applied to genomic data. A real high-dimensional dataset comprising clinical and genetic information of patients with triple-negative breast cancer (TNBC), which is a type of breast cancer with low survival rates due to its aggressive nature, is used to identify variables influencing survival outcomes.

CO266 Room S0.03 RECENT DEVELOPMENTS IN BIOSTATISTICS AND BIOINFORMATICS Chair: Li-Pang Chen

C0389: Polya trees for survival data

Presenter: **Liqun Diao**, University of Waterloo, Canada
Co-authors: Yixing Zhao

Polya trees are commonly used as priors in nonparametric Bayesian analysis. Approaches for utilizing Polya trees to characterize the distribution of time-to-event data are discussed, which may be subjected to different forms of censoring, such as right censoring or interval censoring. Different aspects of Polya trees are covered, including partitions, prior strength, and choices of prior distributions. Comparisons of the proposed methods to existing approaches for estimating survival probabilities are provided in both simulated settings and through applications to real datasets. It is shown that the proposed methods either improve upon or remain competitive with existing nonparametric estimation methods.

C0581: Causal mediation analysis of non-mortality outcomes with follow up truncated by death

Presenter: **An-Shun Tai**, National Cheng Kung University, Taiwan

In the context of mediation analysis, the presence of death-truncated variables poses a challenge as conventional measures fail to accurately assess the role of a mediator in the effect of a treatment on a primary non-mortality outcome. The purpose is to introduce novel estimands survivor natural direct and indirect effects to address this issue. Exchangeability assumptions are employed to mitigate confounding effects, and empirical expressions are derived using information from a pretreatment surrogate variable akin to an instrumental variable. Three estimation approaches, model parameterization, generalized method of moments, and data-adaptive G-computation, are developed and applied using data from a National Emphysema Treatment Trial to illustrate the proposed method.

C0632: Gene expression analysis with SIMEX-trans: Two-phase transfer learning with covariates subject to measurement error

Presenter: **Kaida Cai**, Southeast University, China

The focus is on high-dimensional linear regression in the context of transfer learning, particularly when covariates are subject to measurement errors. A two-phase high-dimensional transfer learning method is proposed using a simulation-extrapolation procedure to improve the performance of target data analysis by leveraging information from auxiliary data sets. Furthermore, the method can adjust estimates to account for the impact of measurement errors when covariate measurement errors are present in both target and auxiliary data sets. In the simulation studies, the method is compared with others that either disregard information from auxiliary data sets or ignore the effect of measurement errors. The results of estimation errors demonstrate that our method outperforms other methods across various scenarios involving different settings of auxiliary data sets and magnitudes of measurement error. The proposed method is applied to gene expression data analysis, showing that it improves gene expression prediction performance in a target tissue by adjusting the effect of measurement errors and integrating information from multiple tissues as auxiliary data sets.

C0766: **Improving prediction with linear regression by shrinking the contributions of the predictors**
Presenter: **Masao Ueki**, Nagasaki University, Japan

An approach is developed to improve prediction with linear regression by shrinking the contributions of the predictors to the predicted value, where the contribution of a predictor is defined as the predictor variable multiplied by the corresponding regression coefficient. To shrink the contribution of a predictor, the soft-thresholding function and its extension induced by the elastic net are applied. As with the familiar variable selection approach, it is possible to eliminate predictor variables if the contribution of a predictor for all subjects is thresholded at zero, but the present approach can result in a more detailed sparse output of the contributions of the predictors for individual subjects' predicted values. Simulation studies confirm that the proposed thresholding improves the linear regression and shows good performance relative to other statistical and machine learning methods. Real data applications, including applications to polygenic risk scores, show an improved predictive performance, as in simulation studies.

CO023 **Room S0.11** STATISTICS IN NEUROSCIENCE II **Chair: Jeff Goldsmith**

C0443: **Algorithmic fairness of models for predicting Alzheimer's disease progression**
Presenter: **Kristin Linn**, University of Pennsylvania, United States

Alzheimer's disease (AD) disproportionately affects marginalized older adults. Machine learning (ML) techniques have the potential to improve early detection of AD. However, ML models may suffer from biases and perpetuate existing disparities. The fairness of three ML models for predicting progression from normal cognition to mild cognitive impairment (MCI) and from MCI to AD is audited. Three common fairness metrics are assessed (equal opportunity, equalized odds, and demographic parity), measured across subgroups defined by gender, ethnicity, and race. Although the three models demonstrated high accuracy in aggregate, all three models failed to satisfy fairness metrics for subgroups defined by ethnicity and race. The models generally satisfied metrics of fairness for gender. Potential implications of the findings are discussed and placed in context with recently published literature on algorithmic fairness.

C0531: **Same data meta analysis: Inference on neuroimaging multiverse data**
Presenter: **Thomas Nichols**, University of Oxford, United Kingdom
Co-authors: Jeremy Lefort-Besnard, Camille Maumet

There is a great diversity of analytical approaches for brain MRI, meaning that there are many possible variations of one result. Integrating a set of such multiverse statistic maps is challenging, as there is strong and possibly complex dependence between the results, violating the independence assumption of traditional meta-analysis. A suite of same-data meta-analysis (SDMA) models are developed that account for dependence among multiple multiverse results from a single dataset. Two main approaches are obtained, one based on the average of inputs ("Stouffer") and one based on generalized least squares (GLS) that optimally combines the correlated data. The validity and accuracy of these models are assessed in a set of simulations as well as on two real-world multiverse outputs originating from the same data: the "NARPS" multiverse analysis and the "HCP Young Adult" multiverse analysis, which generated respectively 70 and 24 different statistical maps. It is shown that all of the methods control false positives, but on real data, the GLS methods can be very unstable if there are complex patterns of correlation among the inputs. These sorts of SDMA approaches are an important tool as more researchers conduct multiverse analyses.

C1157: **Changepoint analysis in a mixed model framework, with applications to fMRI time series**
Presenter: **Mark Fiecas**, University of Minnesota, United States

Motivated by a study on adolescent mental health, a dynamic connectivity analysis is conducted using resting-state functional magnetic resonance imaging (fMRI) data. A dynamic connectivity analysis investigates how the interactions between different regions of the brain, represented by the different dimensions of a multivariate time series, change over time. Changes in the distributional properties of the data can be captured by identifying the changepoints in the time series data. The presence of changepoints in the fMRI data suggests that the connectivity between different regions of the brain changes over time. An overview of changepoint analysis and the utility of dynamic connectivity analysis is given. The novel approach for changepoint analysis that uses a mixed model framework is then described, thereby leveraging the spatial structure of the brain. The mixed model is embedded in a dynamic programming algorithm for detecting multiple changepoints in the fMRI data. The results of the proposed changepoint model in a dynamic connectivity analysis on fMRI are shown on data obtained from female adolescents.

C1156: **Investigating brain functional connectome using regularized blind source separation**
Presenter: **Ying Guo**, Emory University, United States

Brain connectomics has become increasingly important in neuroimaging studies to advance understanding of neural circuits and their association with neurodevelopment, aging and brain disorders. A regularized blind source separation (BSS) framework is presented for reliable mapping of neural circuits underlying static and dynamic brain functional connectome. Using statistical techniques, the proposed framework achieves efficient and reliable mapping of connectivity traits underlying the static and dynamic brain functional connectome, characterizes temporal expressions and interactions of the connectivity traits that contribute to the reconfiguration of the observed dynamic functional connectivity, generates parsimonious and interpretable results in identifying whole-brain connectivity states and identifies connectivity traits associated with demographic attributes and behavioral measures. It is shown that the findings derived from the analyses demonstrate promising reliability and reproductivity across different analytical methods, subject resampling and different studies.

CO256 **Room S0.12** EXTREME VALUE THEORY FOR ENVIRONMENTAL APPLICATIONS **Chair: Juliette Legrand**

C0297: **An extended generalized Pareto regression model for count data**
Presenter: **Touqeer Ahmad**, ENSAI, France
Co-authors: Carlo Gaetan, Philippe Naveau

The statistical modeling of discrete extremes has received less attention than its continuous counterparts in the extreme value theory literature. One approach to the transition from continuous to discrete extremes is the modeling of threshold exceedances of integer random variables by the discrete version of the generalized Pareto distribution. However, the optimal choice of thresholds defining exceedances remains a problematic issue. Moreover, in a regression framework, the treatment of the majority of non-extreme data below the selected threshold is either ignored or separated from the extremes. To tackle these issues, the concept of employing a smooth transition between the bulk and the upper tail of the distribution is expanded on. In the case of zero inflation, models are also developed with an additional parameter. To incorporate possible predictors, the parameters are related to additive smoothed predictors via an appropriate link, as in the generalized additive model framework. A penalized maximum likelihood estimation procedure is implemented. The modeling proposal is illustrated with a real dataset of avalanche activity in the French Alps. With the advantage of bypassing the threshold selection step, the results indicate that the proposed models are more flexible and robust than competing models, such as the negative binomial distribution.

C0474: **A deep geometric approach to modelling multivariate extreme**
Presenter: **Jordan Richards**, School of Mathematics, University of Edinburgh, United Kingdom
Co-authors: Callum Murphy-Barltrop, Reetam Majumder

The geometric representation for multivariate extremes, where data are split into radial and angular components and the radial component is modelled conditionally on the angle, provides an exciting new approach to modelling the extremes of multivariate data. Through a consideration of scaled sample clouds and limit sets, it provides a flexible, semi-parametric model for extremes that connects multiple classical extremal dependence measures; these include the coefficients of tail dependence and asymptotic independence and parameters of the conditional extremes framework. Although the geometric approach is becoming an increasingly popular modelling tool for multivariate extremes, inference with this framework is limited to a low dimensional setting ($d < 4$). The first deep representation is proposed for geometric extremes. By leveraging the predictive power and computational scalability of neural networks, asymptotically-justified yet flexible semi-parametric models are constructed for extremal dependence of high-dimensional data. The efficacy of the deep approach is showcased by modelling the complex extremal dependence between metocean variables sampled from the North Sea.

C0633: **Block maxima modelling in the presence of missing data**
Presenter: **Emma Simpson**, University College London, United Kingdom
Co-authors: Paul Northrop

Modelling block maxima with the generalised extreme value (GEV) distribution is a common method for univariate extremes. One practical challenge in applying this methodology, which is often overlooked, is handling datasets containing missing values. In this case, one cannot be sure whether the true maximum has been observed in each block. If the issue is ignored, this can lead to biased GEV parameter estimates and subsequent underestimation of return levels, which is clearly undesirable when these are to be used for decision-making in practical applications. An approach that is often used to overcome this is to discard blocks where the proportion of missingness exceeds some specified threshold. This means that the chance of having recorded the maximum in each remaining block is reasonably high. While this is a sensible approach, extreme value modelling already comes with an intrinsic lack-of-data issue, so it would generally be preferable not to lose any of the information contained in these discarded blocks. An alternative approach is proposed, where the standard block maxima methodology is extended to overcome missing data issues. The proportion of missing values is explicitly accounted for in each block within the GEV model parameters. The inference is carried out using likelihood-based techniques, with return level estimates and confidence intervals obtained through profiling. The proposed methodology is demonstrated on environmental data.

C0744: **Spatial clustering of multivariate time series based on extremal dependence between sites**
Presenter: **Gwladys Toulemonde**, University of Montpellier; Inria, France
Co-authors: Alexis Boulin, Elena Di Bernardino, Thomas Laloe

Disastrous climate events such as floods or wildfires often occur due to the simultaneous extreme behavior of several interacting processes. The main objective is to develop high-dimensional clustering techniques to handle compound extreme events. The first step consists of developing an extremal dependence summary measure between random vectors in order to quantify extremal dependence between sites on which not only one but several time series could be then considered. The proposed measure becomes a key ingredient in proposing a clustering algorithm for multivariate time series. This method is illustrated by proposing a regionalization task based on gridded data from climate models over Europe. More precisely, based on the ERA5 reanalysis dataset from 1979 to 2022, both daily precipitation sums and daily maximum wind speed data are considered in each pixel.

CO152 Room S0.13 RECENT ADVANCES IN STATISTICS Chair: Trambak Banerjee

C0428: **Harnessing the collective wisdom: Fusion learning using decision sequences from diverse sources**
Presenter: **Trambak Banerjee**, University of Kansas, United States

Learning from the collective wisdom of crowds enhances the transparency of scientific findings by incorporating diverse perspectives into the decision-making process. Synthesizing such collective wisdom is related to the statistical notion of fusion learning from multiple data sources or studies. However, fusing inferences from diverse sources is challenging since cross-source heterogeneity and potential data-sharing complicate statistical inference. An integrative ranking and thresholding (IRT) framework is proposed for fusion learning in multiple testing. IRT operates under the setting where, from each study, a triplet is available: the vector of binary accept-reject decisions on the tested hypotheses, the study-specific false discovery rate (FDR) level and the hypotheses tested by the study. Under this setting, IRT constructs an aggregated, nonparametric, and discriminatory measure of evidence against each null hypotheses, which facilitates ranking the hypotheses in the order of their likelihood of being rejected. IRT guarantees an overall FDR control under arbitrary dependence between the evidence measures as long as the studies control their respective FDR at the desired levels, and a comprehensive numerical study demonstrates that it is a powerful framework for pooling inferences.

C1667: **Sample size requirements to detect an intervention by time interaction in four-level longitudinal CRT**
Presenter: **Priyanka Majumder**, Indian Institute of Science Education and Research Thiruvananthapuram, India
Co-authors: Siuli Mukhopadhyay Siuli Mukhopadhyay

Cluster/group randomized controlled trials (CRTs) have a long history in the study of health sciences. The aim is to design and analyze four-level longitudinal cluster randomized trials. The main interest is to study the difference between treatment groups over time for such a four-level hierarchical data structure. This motivation is based on a real-life study of education-based HIV prevention. Such trials are popular not only for administrative convenience, ethical considerations, and subject compliance but also to help reduce contamination bias. A random intercept mixed model affects linear regression, including a time-by-intervention interaction used for modeling. Closed-form expression of the power function to detect the interaction effect is determined. Sample size equations depend on correlations among schools but not on correlations among classes or students, while the power function depends on the product of the number of units at different levels. The optimal allocation of units under a fixed cost by minimizing the expected standardized variance is also determined and is shown to be independent of correlations among units at any level. Results of detailed simulation studies find the theoretical power estimates based on the derived formulae close to the empirical estimates.

C1668: **A family of nonparametric tests for DMTTF alternatives based on moment inequality**
Presenter: **Shyamal Ghosh**, Indian Institute of Science Education and Research Thiruvananthapuram, India

Based on a moment inequality, a family of test statistics for testing exponentiality against DMTTF alternatives is proposed. The asymptotic distribution of the test statistics is derived under the null and alternative hypothesis, and the consistency of the test is shown by exploiting the U-statistics theory. Comparisons with competing tests are made in terms of Pitman asymptotic relative efficiency (PARE). Additionally, an adapted version of the test under random censorship is explored. The performance of the proposed test has been accessed by means of a simulation study and through application to some real-life data sets.

C0828: **Model multiplicity in policy learning**
Presenter: **Hannah Busshoff**, University of St. Gallen, Switzerland

In policy learning, treatment rules are learned from data that map individual characteristics to treatment recommendations, aiding decision-making in fields like healthcare, education, and public policy. While existing literature derives asymptotic properties of policy learners, it lacks methods to assess the reliability of these rules learned from finite data. Policy rules are derived by choosing the best treatment rule according to training data without accounting for model multiplicity. Model multiplicity refers to the condition where multiple models have similar empirical performance but may imply different individual-level consequences. In deployment, this can cause treatment decisions to be unstable, arbitrary, and unjustified. Formal measures are introduced to quantify the extent of model multiplicity in policy learning. Extending integer programming tools, we analyze decision-relevant model multiplicity, identifying individuals who receive conflicting treatment recommendations from empirically plausible policy rules. The method is demonstrated with synthetic data and applied to two real-world scenarios: allocating individuals to training programs and prescribing tranexamic acid to trauma patients. The applications illustrate the behavior of model multiplicity in two high-stake domains, highlighting its implications for policy reliability.

CO059 Room Safra Lec. Theatre STATISTICAL METHODS FOR FUNCTIONAL AND HIGH-DIMENSIONAL DATA Chair: Eliana Christou

C1069: **Ultra-efficient MCMC for Bayesian longitudinal functional data analysis**
Presenter: **Daniel Kowal**, Cornell University, United States
Co-authors: Thomas Sun

Functional mixed models are widely useful for regression analysis with dependent functional data, including longitudinal functional data with scalar predictors. However, existing algorithms for Bayesian inference with these models only provide either scalable computing or accurate approximations to the posterior distribution, but not both. A new MCMC sampling strategy is introduced for highly efficient and fully Bayesian regression with longitudinal functional data. Using a novel blocking structure paired with an orthogonalized basis reparametrization, the algorithm jointly samples the fixed effects regression functions together with all subject- and replicate-specific random effects functions. Crucially, the joint sampler optimizes sampling efficiency for these key parameters while preserving computational scalability. Perhaps surprisingly, the new MCMC sampling algorithm even surpasses state-of-the-art algorithms for frequentist estimation and variational Bayes approximations for functional mixed models (while also providing accurate posterior uncertainty quantification), and is orders of magnitude faster than existing Gibbs samplers. Simulation studies show improved point estimation and interval coverage in nearly all simulation settings over competing approaches. The method is applied to a large physical activity dataset to study how various demographic and health factors are associated with intraday activity.

C1276: **Dimension reduction for the conditional quantiles of functional data with categorical predictors**
Presenter: **Shanshan Wang**, The University of North Carolina at Charlotte, United States
Co-authors: Eliana Christou, Eftychia Solea, Jun Song

Functional data analysis has received significant attention due to its frequent occurrence in modern applications, such as in the medical field, where electrocardiograms or electroencephalograms can be used for a better understanding of various medical conditions. Due to the infinite-dimensional nature of functional elements, current work focuses on dimension reduction techniques. The focus is to model the conditional quantiles of functional data, noting that existing works are limited to quantitative predictors. Consequently, the first approach is introduced to partial dimension reduction for the conditional quantiles under the presence of both functional and categorical predictors. The proposed algorithm is presented, and the convergence rates of the estimators are derived. Moreover, the finite sample performance of the method is demonstrated using simulation examples and a real data set based on functional magnetic resonance imaging.

C1283: **Sparse independent component analysis with an application to cortical surface fMRI data in autism**
Presenter: **Benjamin Risk**, Emory University, United States
Co-authors: Zihang Wang, Irina Gaynanova, Aleksandr Aravkin

Independent component analysis (ICA) is widely used to estimate spatial resting-state networks and their time courses in neuroimaging studies. It is thought that independent components correspond to sparse patterns of coactivating brain locations. Previous approaches for introducing sparsity to ICA replace the non-smooth objective function with smooth approximations, resulting in components that do not achieve exact zeros. A novel sparse ICA method is proposed that enables sparse estimation of independent source components by solving a non-smooth non-convex optimization problem via the relax-and-split framework. The proposed Sparse ICA method balances statistical independence and sparsity simultaneously and is computationally fast. In simulations, improved estimation accuracy is demonstrated for both source signals and signal time courses compared to existing approaches. The sparse ICA is applied to cortical surface resting-state fMRI in school-aged autistic children. The analysis reveals differences in brain activity between certain regions in autistic children compared to children without autism. Sparse ICA selects coactivating locations, which is argued to be more interpretable than dense components from popular approaches. Sparse ICA is fast and easy to apply to big data.

C1641: **Functional nonlinear mixed-effects modeling of convolved data when direct observations are sparse**
Presenter: **Todd Ogden**, Columbia University, United States
Co-authors: Baoyi Shi, Granville Matheson

In a common PET imaging framework, functional data observed for each of n subjects and k regions are modeled as a convolution between two unknown functions, one of which is common across all regions for each subject. When direct observations from each of these subject-specific common functions are also available, these functions may each be estimated and plugged in prior to model fitting across subjects. However, if such direct data are very sparsely observed, it becomes necessary to model both types of functions simultaneously across subjects and regions. Nonparametric techniques are discussed for model fitting in this situation based on shape constraints that are imposed on the subject-specific common functions. These modeling strategies are illustrated through application to PET neuroreceptor mapping data.

CO070 Room BH (S) 1.01 Lec. Theatre 1 ADVANCES IN MULTIVARIATE TIME SERIES ANALYSIS Chair: Gianluca Cubadda

C0480: **A sequential Monte Carlo approach to estimate noncausal processes**
Presenter: **Francesco Giancaterini**, University of Rome Tor Vergata, Italy
Co-authors: Gianluca Cubadda, Stefano Grassi

A Bayesian estimation approach is introduced for mixed causal and noncausal models using sequential Monte Carlo (SMC) methods in univariate and multivariate contexts. The SMC method allows the estimation of the investigated process by considering different assumptions about the distributions of the error term. Consequently, the SMC approach facilitates the comparison of marginal data densities under different assumptions, helping to identify the error-term assumption that best fits the data. Furthermore, SMC offers extensive parallelization possibilities, significantly reducing estimation time and mitigating the risk of becoming trapped in local minima. Simulation studies demonstrate the strong ability of the algorithm to correctly identify the process and provide accurate estimates.

C0503: **Reduced-rank matrix autoregressive models: A medium N approach**
Presenter: **Ivan Ricardo**, Maastricht University, Netherlands

Reduced-rank regressions are powerful tools used to identify co-movements within economic time series. However, this task becomes challenging when matrix-valued time series are observed, where each dimension may have a different co-movement structure. Reduced-rank regressions are proposed with a tensor structure for the coefficient matrix to provide new insights into co-movements within and between the dimensions of matrix-valued time series. Moreover, the co-movement structures are related to two commonly used reduced-rank models, namely the serial correlation common feature and the index model. Two empirical applications involving U.S. states and economic indicators for the Eurozone and North American countries illustrate how the new tools identify co-movements

C1086: **Green bubbles: A noncausal approach**

Presenter: **Alain Hecq**, Maastricht University, Netherlands

The phenomenon of the "green bubble" has drawn attention in the field of renewable energy investment, raising concerns about potential economic and environmental implications. An investigation into mixed causal and noncausal models applied to the Renixx green bubble indicator is presented, offering an innovative approach to analyzing such phenomena. Departing from traditional definitions of bubbles, a perspective is adopted in which the Renixx price index is viewed as following a strictly stationary process, with the bubble considered an inherent component of its dynamics. Through the exploration of causal-noncausal autoregressive processes, the aim is to uncover insights into the dynamics of green bubbles and their broader implications. The findings contribute to the understanding of financial and environmental sustainability in the context of investments in renewable energy.

C1116: **Exploring noncausal and noninvertible ARMA-GARCH dynamics in the cryptocurrency market**

Presenter: **Daniel Velasquez-Gaviria**, Maastricht University, Netherlands

Co-authors: Alain Hecq

The prices of Bitcoin and Ethereum exhibit non-stationarity, marked by frequent episodes of local explosions that abruptly collapse. This behavior resembles the well-known formation of financial bubbles, motivating the use of the mixed causal-noncausal and invertible-non-invertible autoregressive moving average (MARMA) model, which can replicate these characteristics. Additionally, conditional volatility is evident, where episodes of large variations are followed by further large variations and episodes of small variations are followed by further small variations. The estimation and identification of the new MARMA-GARCH model is proposed. An estimation in both the frequency domain and the time domain is suggested and promising techniques are introduced to remove the price trend and estimate the model in the cyclical component. The results indicate that Bitcoin and Ethereum prices exhibit noncausal behaviors, along with significant conditional volatility.

CO369 Room BH (SE) 1.01 BAYESIAN HIGH-DIMENSIONAL REGRESSION AND MODEL SELECTION Chair: Somak Dutta

C1206: **Informed MCMC for Bayesian variable selection**

Presenter: **Vivekananda Roy**, Iowa State University, United States

A Riemannian geometric framework for Markov chain Monte Carlo (MCMC) is developed where using the Fisher-Rao metric on the manifold of probability density functions (PDFs) informed proposal densities for Metropolis-Hastings (MH) algorithms are constructed. The square-root representation of PDFs is exploited, under which the Fisher-Rao metric boils down to the standard L2 metric, resulting in a straightforward implementation of the proposed geometric MCMC methodology. Unlike the random walk MH that blindly proposes a candidate state using no information about the target, the geometric MH algorithms effectively move an uninformed base density (e.g., a random walk proposal density) towards different global/local approximations of the target density. This general geometric framework is used to construct fast mixing and scalable MCMC algorithms for performing Bayesian variable selection based on a hierarchical Gaussian linear model with popular spike and slab priors. The superiority of the geometric MH algorithm over other MCMC schemes is demonstrated using extensive ultra-high dimensional simulation examples, as well as a real dataset from a genome-wide association study (GWAS) with close to a million markers.

C1207: **Reproducible Bayesian model selection and high-dimensional regression**

Presenter: **Jonathan Huggins**, Boston University, United States

If slightly changing a model specification or including more data results in contradictory inferences, then the validity of any conclusions drawn from such inferences is put in doubt: they are not, in a statistical sense, reproducible. Motivated by examples ranging from phylogenetic tree reconstruction to crime rate prediction, the aim is to discuss how model misspecification can result in standard Bayesian inference, leading to such non-reproducible results. An easy-to-implement solution, the bagged posterior, is also outlined.

C1455: **Non-asymptotic Laplace approximation under model misspecification**

Presenter: **Anirban Bhattacharya**, Texas AM University, United States

Co-authors: Debdeep Pati

Non-asymptotic two-sided bounds are presented to the log-marginal likelihood in Bayesian inference. The classical Laplace approximation is recovered as the leading term. The derivation permits model misspecification and allows the parameter dimension to grow with the sample size. No assumptions are made about the asymptotic shape of the posterior, and instead, certain regularity conditions on the likelihood ratio and that the posterior is sufficiently concentrated. The derived bounds are envisioned to be widely applicable in establishing model selection consistency of Bayesian procedures in non-conjugate settings, especially when the true model potentially lies outside the class of candidate models considered.

C1464: **Bayesian variable selection for multi-layer biological data**

Presenter: **Hao Cheng**, University of California Davis, United States

Advances in high-throughput sequencing technology have generated an increasing volume and diversity of multi-omics data that complement genomics. The effects of genetic variants (e.g., SNPs) on phenotypes can be mediated by multiple layers of molecular variations through mechanisms such as regulatory cascades from the genome to the transcriptome and proteome. These molecular variations serve as measurable intermediates between DNA and phenotype and are partially heritable across generations. In most models incorporating intermediate molecular variations for complex trait prediction, these variations are typically treated as independent variables alongside SNPs or as responses in addition to empirical complex traits in mixed models. This approach overlooks the sequential relationships between genetic variants, intermediate molecular variations, and complex traits. A Bayesian variable selection method is developed that extends mixed models into multilayer neural networks to capture the nonlinear relationships between genotypes and phenotypes, improving prediction and inference for complex traits. The neural network is further developed, named NNMM (a mixed-effect neural network), to incorporate intermediate omics features into the middle layers of the network. This enables mechanistic modeling of the regulatory pathways from genotypes, through intermediate omics features, to phenotypes.

CO107 Room BH (SE) 1.02 BAYESIAN SEMI- AND NON-PARAMETRIC METHODS II Chair: Andres Barrientos

C0903: **Bayesian mapping of mortality clusters**

Presenter: **Andrea Sottosanti**, University of Padova, Italy

Disease mapping techniques study the distribution of various disease outcomes within a territory, aiming to identify areas with unexpected mortality rate changes, analyze connections across diseases, and divide territories into clusters based on disease incidence or mortality levels. The focus is on detecting spatial mortality clusters that occur when neighbouring areas within a territory exhibit similar mortality levels due to multiple diseases. Identifying both spatial boundaries and responsible diseases is crucial, yet existing methods fail to address this dual problem effectively. To overcome these limitations, Perla is proposed, a multivariate Bayesian model that clusters the areas in a territory according to the observed mortality rates of multiple causes of death, also exploiting the information of external covariates. Perla incorporates the spatial structure into clustering probabilities using the multinomial stick-breaking, and it exploits suitable global-local shrinkage priors to ensure that the detection of clusters is driven by concrete differences across mortality levels while excluding spurious differences. An MCMC algorithm is developed for posterior inference with closed-form Gibbs sampling for nearly all model parameters, requiring minimal tuning. The effectiveness of the methodology is demonstrated by analyzing mortality levels in two distinct territories: the Padua province in northeastern Italy and the counties on the east coast of the U.S.

C1063: **Bayesian monotone single-index quantile regression model with bounded response and misaligned functional covariates**
Presenter: **Debajyoti Sinha**, Florida State University, United States

The main goal is to understand how existing scalar variables, as well as multiple functional covariates measuring neural response to rewards, are associated with future adolescent depression. Unlike previous studies using simple linear regression to index all covariates, a novel Bayesian quantile regression model is proposed using a single-index summary of all scalar and functional covariates along with an associated monotone link function to accommodate unknown functional forms as well as interactions among the covariates. Compared to existing methods, the new method also addresses the following practical challenges: an index with a clinical interpretation similar to a linear model, a fitted value of the pre-specified quantile within the same bounds as the response, and accommodation of the uncertainty in the registration/alignment of the observed functional covariates within the data analysis. In the simulation, the new method outperforms existing unrestricted single-index-based models in the presence of both scalar and even pre-registered functional covariates. The practical advantages and implications of the method are illustrated by analyzing a large existing adolescent depression study and, in the process, developing a new statistically principled summary of the functional covariates measuring neural response to rewards.

C1109: **Model-based clustering of categorical data based on the Hamming distance**
Presenter: **Raffaele Argiento**, Universita degli Studi di Bergamo, Italy
Co-authors: Lucia Paci, Edoardo Filippi-Mazzola

A model-based approach is proposed for clustering categorical data without a natural ordering. The proposed method leverages the Hamming distance to create a family of probability mass functions for modelling the data. These functions serve as kernels within a finite mixture model with an undetermined number of components. Conjugate Bayesian inference has been developed for the parameters of the Hamming distribution model. The mixture is situated within a Bayesian nonparametric framework, and a trans-dimensional blocked Gibbs sampler is introduced to facilitate comprehensive Bayesian inference on the number of clusters, their structure, and the group-specific parameters. This approach simplifies computation compared to traditional reversible jump algorithms. The proposed model includes a parsimonious latent class model as a special case when the number of components is predetermined. Model performance is evaluated through simulation studies and benchmark datasets, demonstrating improvements in clustering accuracy over existing methods.

C1692: **Posterior uncertainty quantification in neural networks using data augmentation**
Presenter: **Sinead Williamson**, Apple, United States

The focus is on the problem of uncertainty quantification in deep learning through a predictive framework, which captures uncertainty in model parameters by specifying our assumptions about the predictive distribution of unseen future data. Under this view, we show that deep ensembling is a fundamentally mis-specified model class, since it assumes that future data are supported only by existing observations- a situation rarely encountered in practice. To address this limitation, we propose MixupMP, a method that constructs a more realistic predictive distribution using popular data augmentation techniques. MixupMP operates as a drop-in replacement for deep ensembles, where each ensemble member is trained on a random simulation from this predictive distribution. Grounded in the recently proposed framework of Martingale posteriors, MixupMP returns samples from an implicitly defined Bayesian posterior. Our empirical analysis showcases that MixupMP achieves superior predictive performance and uncertainty quantification on various image classification datasets, when compared with existing Bayesian and non-Bayesian approaches.

CO296 Room BH (SE) 1.05 CHANGE-POINT DETECTION FOR HIGH-DIMENSIONAL OR NON-EUCLIDEAN DATA Chair: Lynna Chu

C0236: **Change point inference in high-dimensional regression models under temporal dependence**
Presenter: **Daren Wang**, Notre Dame, United States
Co-authors: Haotian Xu, Zifeng Zhao, Yi Yu

The focus is on the limiting distributions of change point estimators in a high-dimensional linear regression time series context, where a regression object is observed at every time point. At unknown time points, called change points, the regression coefficients change, with the jump sizes measured in L_2-norm. Limiting distributions of the change point estimators are provided in the regimes where the minimal jump size vanishes and where it remains constant. Both the covariate and noise sequences are temporally dependent in the functional dependence framework, which is the first time seen in the change point inference literature. A block-type long-run variance estimator is shown to be consistent under functional dependence, which facilitates the practical implementation of the derived limiting distributions. A few important byproducts of the analysis are also presented, which are of their own interest. These include a novel variant of the dynamic programming algorithm to boost computational efficiency, consistent change point localization rates under temporal dependence and a new Bernstein inequality for data possessing functional dependence. Extensive numerical results are provided to support the theoretical results.

C0465: **Practical and powerful kernel-based change-point detection**
Presenter: **Hoseung Song**, KAIST, Korea, South

Change-point analysis plays a significant role in various fields to reveal discrepancies in distribution in a sequence of observations. While a number of algorithms have been proposed for high-dimensional data, kernel-based methods have not been well explored due to difficulties in controlling false discoveries and mediocre performance. A new kernel-based framework is proposed that makes use of an important pattern of data in high dimensions to boost power. Analytic approximations of the significance of the new statistics are derived, and fast tests based on the asymptotic results are proposed, offering easy off-the-shelf tools for large datasets. The new tests show superior performance for a wide range of alternatives when compared with other state-of-the-art methods. These new approaches are illustrated through an analysis of phone-call network data. All proposed methods are implemented in an R package kerSeg.

C0622: **High-dimensional change-point detection using generalized homogeneity metrics**
Presenter: **Runmin Wang**, Texas A&M University, United States
Co-authors: Xianyang Zhang, Shubhadeep Chakraborty

The purpose is to study the problem of detecting abrupt changes in the data-generating distributions of a sequence of high-dimensional observations beyond the first two moments. This problem has remained substantially less explored in the existing literature, especially in the high-dimensional context, compared to detecting changes in the mean or the covariance structure. A nonparametric methodology is developed to (i) test the existence of a change-point and (ii) identify the change-point locations in an independent sequence of high-dimensional observations. The approach rests upon recent nonparametric tests for the homogeneity of two high-dimensional distributions. A single change-point test statistic is constructed based on a cumulative sum process in an embedded Hilbert space. Its limiting null distribution is derived, and the asymptotic consistency is presented under the high dimension medium sample size framework. The statistics are also combined with wild binary segmentation to recursively estimate and test for multiple change-point locations. The superior performance of the methodology compared to other existing procedures is illustrated via extensive simulation studies and the application to the stock return data observed during the period of the global financial crisis in the United States.

C1077: **Integrating privacy enhancements with dynamic community detection**

Presenter: **Liyan Xie**, University of Minnesota, United States

In the evolving landscape of online communities, safeguarding user privacy while accurately detecting dynamic changes presents a critical challenge. The private online community detection problem is studied by integrating edge differential privacy (DP) within the framework of a censored block model (CBM). The fundamental tradeoffs between the privacy budget, detection performance, and exact community recovery of community labels are explored. The proposed algorithm can identify changes in the community structure while maintaining user privacy. A new information-theoretic lower bound is also established on the delay in detecting community changes privately.

CO380 Room BH (SE) 1.06 MODERN SEASONALITY AND RISK ANALYSIS Chair: Yushu Li

C1016: **Modern portfolio theory with seasonal assets**

Presenter: **Sondre Hoelleland**, Norwegian School of Economics, Norway

Co-authors: Haakon Otneim, Geir Drage Berentsen, Konstantinos Fokianos

The main application of modern portfolio theory (MPT) is in financial portfolio optimization. For a given expected return, the portfolio variance is minimized. The finance literature has moved on to using other risk measures, but the MPT has found other application fields. Recent applications include renewable energy planning, such as citing wind farms, optimal wind-solar energy mixes or cross-border energy resource allocation. Most renewable energy sources have strong seasonal dynamics that the traditional MPT framework does not consider. The consequences of applying the standard procedure on assets are analyzed with underlying seasonal dynamics. A modified formulation of the classical MPT is proposed, leaving the investor in complete control over how seasonality should be handled in a given practical problem. The suggested approach is demonstrated using empirical examples from renewable energy resource allocation problems.

C1361: **Particle Markov chain Monte Carlo for parameter estimation in volatility models**

Presenter: **Eivind Lamo**, University of Bergen, Norway

With the continuous increase in computational power, sequential Monte Carlo methods have emerged as an efficient technique for estimating unknown data in a world consisting of nonlinearity and non-Gaussianity. A theoretical foundation is built with the help of Bayesian statistics that can be applied to numerous real-world problems. The interest is in solving the problem of estimating an unknown signal process given certain observations, where both processes are modelled as Markovian, nonlinear, non-Gaussian state-space models. In particular, the attempt is to estimate the unobserved volatility dynamics for the S&P 500 index using observed returns and a slight modification of Heston's stochastic volatility model. This will be done using the sequential importance resampling filter, which is also combined with the Markov chain Monte Carlo for parameter estimation. The overall goal is to propose another alternative to Heston's model, by investigating how well the model responds to measuring volatility when including data from the financial crisis of 2007-2008.

C1490: **Forecasting jointly value at risk and expected shortfall for energy commodities using quantile regression**

Presenter: **Sjur Westgaard**, Norwegian University of Science and Technology, Norway

Expected shortfall (ES) offers advantages over value at risk (VaR) by addressing VaRs limitations in capturing extreme tail risks. While VaR estimates the maximum loss at a given confidence level, ES provides the average loss beyond this threshold, offering a more comprehensive measure of extreme risks. Consequently, ES is increasingly adopted in risk assessment frameworks, with the Basel Committee on Banking Supervision, recommending a shift from VaR to ES. Scenario analysis and stress testing are also crucial for assessing the impact of extreme events on financial institutions. Quantile regression models estimating VaR and ES jointly has been developed by past studies. The framework also facilitates scenario analysis and stress testing by investigating how changes in the risk factors directly influence VaR and ES at different quantiles. The R package "esreg" implements these methods, and "esback" offers joint VaR and ES prediction tests. The purpose is to explore these methods in estimating VaR and ES for energy commodities like oil, natural gas, electricity, and coal, and examine the influence of risk factors on tail risks.

C1476: **Time-varying multi-seasonal AR models**

Presenter: **Mattias Villani**, Stockholm University, Sweden

A seasonal AR model with time-varying parameter processes in both the regular and seasonal parameters is presented. The model is parameterized to guarantee non-explosive behavior at every time point and can accommodate multiple seasonal periods. Time evolution is modelled by dynamic shrinkage processes to allow for longer periods of essentially constant parameters and rapid jumps. A robust, fast and accurate approximate sampler based on the extended Kalman filter is proposed and compared to a particle MCMC sampler. The properties of the model are compared to several benchmark models on simulated data. An application to more than a century of monthly US industrial production data shows interesting changes in seasonality over time, particularly during the Great Depression and the recent COVID-19 pandemic.

CO347 Room BH (S) 2.01 ECONOMETRICS APPLIED TO FINANCE AND INSURANCE Chair: Masayuki Hirukawa

C0207: **Nonparametric estimation of splicing points in cost distributions through a transformation-based approach**

Presenter: **Benedikt Funke**, Cologne University of Applied Sciences, Germany

Co-authors: Masayuki Hirukawa

The tail of a distribution beyond some threshold is of importance and interest in academics and industries. The aim is to investigate the nonparametric estimation of a threshold in distributions of nonnegative cost variables such as incomes or insurance payments. It starts by interpreting the threshold as a jump location or splicing point in a distribution. Since the threshold typically lies in the right-tail region, it is often difficult to estimate it due to a small jump size and/or sparseness of the data. Then, the proposal is to transform the original nonnegative observations onto the unit interval and detect the threshold in the transformed scale using the asymmetric beta kernel. The data transformation makes the threshold estimation easier because the jump size is magnified, and two adjacent observations become closer to the unit interval. It is demonstrated that the threshold estimator is consistent with a faster convergence rate than the parametric one and asymptotically normal when suitably implemented. Monte Carlo simulations and real data examples illustrate attractive properties and practical relevance of the proposal in several different use cases.

C0223: **Integrating climate risks and nonlinear dependencies into solvency assessment for non-life insurers**

Presenter: **Onur Oezdil**, Friedrich-Alexander-Universitaet Erlangen-Nuernberg, Germany

In light of increasing physical and transitional climate risks, understanding their combined impact on assets and liabilities is crucial for the financial stability of non-life insurers, as they are significantly affected by both types of risks. The aim is to analyze how climate risks influence the solvency requirements of non-life insurers, where both physical and transitional risks, along with their potential interdependencies, are modeled using copulas with tail dependence. Rather than focusing on the direct effects on solvency capital requirements, the emphasis is on minimum risk and return requirements, also referred to as solvency lines, for brown, green, and other types of investments. Given that asset allocations are more adaptable and accessible in response to short-term climate shocks compared to equity or liability structures, they are anticipated to be the primary focus of management actions for maintaining solvency. The analysis of different climate scenarios reveals their potential effects on the permissible investment portfolios of insurers over several years. Specifically, insurers might need to achieve an additional annual return of approximately 6% for a one-year time horizon and 3.2% for a ten-year period, whereby these requirements vary significantly based on the composition of the assets and their interdependencies.

C0239: **Detecting changes in the strength of dependence between financial data sets**
Presenter: **Alexander Schnurr**, University Siegen, Germany

The concept of ordinal pattern dependence is recalled between time series and shows in an explorative study that both types of this dependence show up in real world financial data. The classical way to capture the leverage effect in models for stock markets is to assume a negative correlation between the two datasets, which is constant in time. However, there is strong evidence that this effect is not constant but evolves over time. It seems that there are periods where the effect is weaker, and sometimes, it even seems to be turned around. Taking these empirical findings into account, more sophisticated models were suggested. The correlation structure was modeled by a deterministic function, was made state space dependent or even modeled itself by a stochastic process. Instead of proposing more complicated models, a rather simple approach is introduced to analyze whether there is a dependence structure between two datasets: In order to capture the zigzag of datasets, so-called ordinal patterns are used. From this point of view, the two datasets are compared. On some occasions, as an example, the S&P 500 and the VIX are considered; a dependence structure of this kind seems to be more likely to be found in real data than the dependence modeled by the classical approach via correlation.

C0315: **Time-varying coefficient regression models: Estimation and prediction by local linear smoothers using asymmetric kernels**
Presenter: **Masayuki Hirukawa**, Ryukoku University, Japan

Trending time-varying coefficient regression models are investigated. Time-varying coefficients are estimated nonparametrically by local linear ("LL") regression smoothing. Because the domain of varying coefficients is [0,1], nonstandard, asymmetric kernels ("AKs") that are free of boundary bias are employed. Among all such kernels, a particular focus is on the beta and gamma kernels due to their popularity in empirical studies in economics and finance. Convergence properties of AK-LL estimators for varying coefficients at a fixed design point and in the vicinity of 1, including their bias and variance approximations and asymptotic normality, are explored. Their finite-sample properties, as well as the effectiveness of an implementation method, are examined via Monte Carlo simulations. Implications for prediction are also considered, in combination with long-run variance estimation for asymptotic variances of AK-LL estimators.

CO151 Room BH (S) 2.02 ADVANCES IN STATISTICAL MODELLING FOR ECONOMICS AND FINANCE Chair: Silvia Angela Osmetti

C0757: **Nonparametric modelling of EUA market returns, volatility, and financial market links: A data-driven approach**
Presenter: **Cristiano Salvagnin**, University of Brescia, Italy
Co-authors: Aldo Glielmo, Maria Elena De Giuli, Antonietta Mira

An exhaustive analysis of volatility dynamics in the European Union Emissions Trading System (EU ETS) market is conducted using a nonparametric framework enhanced by differentiable information imbalance (DII). By incorporating time-varying information imbalances, the primary drivers of volatility are revealed, and complex patterns such as volatility clusters and dependence structures are captured. The approach leverages DII's flexibility and adaptability to effectively model the intricate relationships in financial data. Applying the methodology to the financial returns and realized volatility of the EU ETS market, key volatility determinants are identified, including commodities, energy indices, exchange rates, uncertainty indicators, and macroeconomic factors. Findings provide deep insights into the nature of volatility within the EU ETS market, offering actionable intelligence for market participants, policymakers, and researchers. Specifically, it is found that the EU ETS market exhibits a strong one-way connectedness with commodities in terms of financial returns. When analyzing realized volatility, energy indices display significant connectedness and spillover effects, impacting commodities with high relevance. The newly developed methodology, focused on data dispersion into space and data manifold, provides novel insights and advanced tools for understanding non-linear relationships in a nonparametric framework.

C0478: **Social media information to forecast Bitcoin value: A comparison of vines and graphical models**
Presenter: **Lorenzo Merli**, University of Pavia, Italy
Co-authors: Claudia Tarantola, Luciana Dalla Valle, Silvia Angela Osmetti

The aim is to enhance Bitcoin price forecasts by leveraging graphical models and vine copulas. By integrating daily Bitcoin prices with Google Trends data, Twitter activity, and sentiment analysis using Bing and Afinn lexicons, the complex relationships within Bitcoin trends are captured. One hundred fourteen (114) daily observations from February to May 2021 are utilized. Mixed graphical models (MGM) and vector autoregressive (VAR) models forecast Bitcoin prices, while ARIMA-GARCH and gamlss models extract residuals for vine copula implementation. Vine models predict Bitcoin prices using a rolling window method. Comparing forecasts with observed data highlights model accuracy, providing a comprehensive view of Bitcoin market dynamics and public sentiment.

C0861: **Ordered response models for cyber risk assessment**
Presenter: **Silvia Facchinetti**, Universita Cattolica Del Sacro Cuore Di Milano, Italy
Co-authors: Silvia Angela Osmetti, Claudia Tarantola

Evaluating the risk of cyber-attacks is essential for companies. There is a growing need to develop and implement effective strategies for cyber security, data security, and privacy protection. With the rise in cyber threats, assessing the risk of a successful attack is increasingly important for companies and their customers. While quantitative loss data are seldom available, experts can provide qualitative evaluations of attack severity on an ordinal scale. Hence, the ordered response model, particularly the cumulative link model, is suitable for analyzing cyber risk. This model explains the experts' assessments of the severity of a cyber-attack based on a set of explanatory variables describing the characteristics of the attack under consideration, including measures of the attack's impact diffusion through a network structure. Additionally, a detailed analysis of a real dataset is offered, documenting major cyber-attacks worldwide from 2017-2018.

C1126: **CEOs on social media and stock market predictability**
Presenter: **Suyong Song**, University of Iowa, United States
Co-authors: Kang-Pyo Lee

Machine learning techniques are applied to high-dimensional social media data from CEO postings, and they have been shown to be useful in predicting stock market indicators. We create a large, unique sample of CEO users on Twitter, and construct hashtag and sentiment time series. Findings confirm that the select list of hashtags and sentiments have predictive power on the stock return, trading volume, volatility, and stock price direction. We also find that the predictive power of CEO sentiments still stands after controlling for well-known macroeconomic and financial variables.

CO327 Room BH (S) 2.03 TIME SERIES ECONOMETRICS Chair: Josu Arteche

C0291: **Modeling and forecasting the long memory of cyclical trends in paleoclimate data**
Presenter: **Philipp Sibbertsen**, University of Hannover, Germany
Co-authors: Tomas del Barrio Castro, Alvaro Escribano

The relevant cycles are identified and estimated in paleoclimate data of earth temperature, ice volume and CO2. Cyclical cointegration analysis is used to connect these cycles to the earth's eccentricity and obliquity and to see that the earth's surface temperature and ice volume are closely connected. These findings are used to build a forecasting model that includes the cyclical component as well as the relevant earth and climate variables, which outperforms models by ignoring the cyclical behavior of the data. The turning points can be especially predicted accurately using the proposed approach. Out-of-sample forecasts for the turning points of earth temperature, ice volume and CO2 are derived.

C0345: **Long memory in the marginalized time series of a VAR revisited**
Presenter: **Tomas del Barrio Castro**, University of the Balearic Islands, Spain
Co-authors: Philipp Sibbertsen, Andreu Sanso

The purpose is to demonstrate analytically and through Monte Carlo results that the long memory observed in the marginalized univariate time series of a VAR(1), as defined by prior studies, depends not only on the value assigned to the main diagonal of the Toeplitz matrix associated with the parameters of the VAR(1) but also on the number of time series in the VAR. Using the damped trend representation proposed by a recent study, it is shown that as the number of time series in the VAR(1) increases, the long memory in the marginalized univariate time series decreases. The analysis is also extended to VAR(2) models, allowing for long memory associated with harmonic frequencies in the marginalized time series. Finally, the marginalized time series in these VAR models are pointed out to be cointegrated with each other, as the long memory behavior is governed by the damped trend.

C0518: **Estimation of time-varying long memory series**
Presenter: **Josu Arteche**, University of the Basque Country, Spain
Co-authors: Luis Filipe Martins

The memory parameter is usually assumed to be constant in traditional long memory time series. This restriction is relaxed by considering the memory a time-varying function that depends on a finite number of parameters. A time-varying local Whittle estimator of these parameters, and hence of the memory function, is proposed. Its consistency and asymptotic normality are shown for locally stationary and locally non-stationary long memory processes, where the spectral behavior is restricted only at frequencies close to the origin. Its good finite sample performance is shown in a Monte Carlo exercise and in two empirical applications, highlighting its benefits over the fully parametric Whittle estimator proposed in a past study. Standard inference techniques for the constancy of the memory are also proposed based on this estimator.

C0651: **Time domain estimation of non-fundamental ARMA models in the presence of heteroskedasticity of unknown**
Presenter: **Carlos Velasco**, Universidad Carlos III de Madrid, Spain
Co-authors: Ignacio Lobato

Time domain estimation of possibly non-fundamental (that is, non-causal and/or non-invertible) non-Gaussian linear ARMA models with martingale difference innovations that may display conditional heteroskedasticity of unknown form is considered. Instead of explicitly parametrizing the underlying volatility process (the higher order dependence) and employing maximum likelihood procedures, the unpredictability of the true model innovations is exploited, and a time domain minimum distance estimator is proposed based on innovations predictability using second and third powers of past innovations. Using the proposed estimator, which is consistent and asymptotically normal, an ARMA(1,1) model is fit to the squares of exchange rate returns, and evidence of non-invertibility in three out of seven cases is found.

CO217 Room BH (S) 2.05 ECONOMETRICS OF ART MARKETS Chair: Douglas Hodgson

C0242: **Career profiles of design quality for golf course architects**
Presenter: **Douglas Hodgson**, UQAM, Canada
Co-authors: Daniel Ackerberg

The effects of age or career experience on productivity, as measured by the creative quality of the work for workers in creative fields, has been of some interest to researchers in cultural economics. One line of research has focused on successive generations of painters in environments where the field has transitioned from an emphasis on artisanal craftsmanship to one of avant-garde originality as the principal criterion of judgments of quality, which are measured by auction prices of paintings. For various possible reasons, career creativity profiles in such environments have often been found to shift to the left so that later generations have creativity peaks that occur earlier in life than their predecessors. One could argue that a similar evolution in standards occurred during the twentieth century in the field of golf course architecture and that one might observe similar intra-generational shifts in average career profiles. This hypothesis is evaluated with two novel data sets and corresponding econometric methodologies: First, numerical rankings of golf course quality obtained from a popular guide to international golf courses for courses designed by a set of major golf course architects with birth dates covering a range of over 150 years, and second, magazine rankings of the top 100 golf courses in the United States. Surprisingly, no evidence is found of the hypothesized shift, and if anything, a reverse shift may be present. Possible explanations are suggested.

C0245: **Estimating the value of psychic income for artists and other workers**
Presenter: **Joanna Woronkowicz**, Indiana University, United States
Co-authors: Douglas Noonan

The notion of the gig worker in the modern economy, where working arrangements are temporary or project-based, has brought up a slew of policy questions related to whether the nonpecuniary value these workers might receive from engaging in gig work makes up the difference between what they don't receive in terms of wages and other pecuniary benefits. Employment data is used on workers in the United States from the 2018-2022 American Community Survey collected by the U.S. Census Bureau to estimate the value of psychic income for artists and other workers, including gig workers. Since accounting for the implicit prices of job characteristics across occupations can be essential to fully understand the compensation and selective pressures associated with the workforce, a selection model is used to separate the selective pressures from observable and unobservable worker characteristics. The results show that, among artist workers, the average estimated wage differential is not consistently positive. By contrast, primary school teachers experience a consistently positive amount of psychic income. In light of the nature of the sample, identification strategy, and market equilibrium conditions, the implications of these results are discussed in relation to gig-style work, especially those relevant to creative occupations.

C0459: **A comparison of the career profiles of Australian Impressionist artists**
Presenter: **Bronwyn Coate**, RMIT University, Australia
Co-authors: Douglas Hodgson

The existence of artistic movements in which small groupings of artists share aesthetic or programmatic similarities and utilise group association to further their collective programme and individual careers and creative trajectories is well established. Yet, despite widespread acknowledgement of the importance of artistic movements in the context of understanding the canon of art, the contributions of individual artists in first forming artistic movements are less explored. Regression modelling and archival methods are used to examine the careers of artists associated with the Australian Impressionist art movement. Of key interest is the aim to understand how artists' centrality to the formation of the Australian Impressionist movement impacts their career trajectories. To measure artists' centrality within the formation of the movement, evidence is drawn upon from the catalogue for the 9 x 5 exhibition held in Melbourne in 1889, which, despite the sharp criticism it received at the time, is associated with the start of Impressionism in Australia. The analysis also draws on data about artists' representation in the collections of major public galleries across Australia to provide further evidence on artist reputation.

C1002: **The "Motherhood Penalty" in artistic production: Historical evidence from American authors, 1800-1999**
Presenter: **Christiane Hellmanzik**, Technical University of Dortmund, Germany
Co-authors: Lukas Kuld, Sara Mitchell

Historical data is utilised to explore the existence of a child penalty in authorship and to determine whether female authors were disproportionately affected. A novel dataset of 472 eminent American writers (born 1800-1949) that includes yearly residence, the year in which their children were born, and data on career success (including publications, critical acclaim, and market success data) is used. The productivity of male and female authors following the birth of a child is investigated. Significantly lower publication rates for female authors are found for the first ten years following the birth of a child. A similar reduction in productivity was not observed for male authors. These findings are confirmed using the general fertility rate, age, and gender as an IV for the probability of motherhood. The long-run trend of this child penalty is also explored from a time in which women's rights were severely restricted through the expansion of women's rights through the mid-20th century. Furthermore, how this motherhood penalty evolved post-WWII when there was a divergence in gender representation in the labor force, a baby boom, and rigid gender roles is explicitly investigated.

CO012 Room BH (SE) 2.01 HIGH-DIMENSIONAL DATA Chair: Nicola Loperfido

C0240: **Precision nosology for mental health research**
Presenter: **Thaddeus Tarpey**, New York University, United States

Nosology is the branch of medical science dealing with the classification of diseases. Nosological research has focused on prototypical approaches postulating the existence of distinct disease categories and dimensional approaches where disease severity varies along a continuum with no clear demarcating disease boundaries. In some mental health diseases, such as depression and ADHD, disease categories have become more and more subdivided across the distribution of diagnostic features to accommodate heterogeneity within reified disease categories. This refining of psychopathology points to the need for precision nosology to individualize diagnoses and treatment based on patient-specific features across data modalities. This is clearly a problem where statistical modeling can play an important role. Common statistical approaches have used tools such as factor analysis and clustering. The use of independent component analysis is proposed to obtain flexible nonparametric density estimates in high-dimensional feature spaces that can accommodate skewing effects due to psychopathology. This approach allows for a nosologically-based projection pursuit modeling to help discover lower-dimensional subspaces to guide the discovery of disease heterogeneity. Approaches that can incorporate data from multiple modalities will also be discussed.

C0243: **Semi-supervised sparse Gaussian classification: Provable benefits of unlabeled data**
Presenter: **Boaz Nadler**, Weizmann Institute of Science, Israel
Co-authors: Eyar Azar

The premise of semi-supervised learning (SSL) is that combining labeled and unlabeled data yields significantly more accurate models. Despite empirical successes, the theoretical understanding of SSL is still far from complete. SSL is studied for high dimensional sparse Gaussian classification. A key task in constructing an accurate classifier is feature selection, which detects the few variables that separate the two classes. For this SSL setting, information-theoretic lower bounds are analyzed for accurate feature selection as well as computational lower bounds, assuming the low-degree likelihood hardness conjecture. The key contribution is the identification of a regime in the problem parameters (dimension, sparsity, number of labeled and unlabeled samples) where SSL is guaranteed to be advantageous for classification. Specifically, there is a regime where it is possible to construct an accurate SSL classifier in polynomial time. However, computationally efficient supervised or unsupervised learning schemes that separately use only the labeled or unlabeled data would fail. The provable benefits of combining labeled and unlabeled data are highlighted for classification and feature selection in high dimensions. Simulations are presented that complement the theoretical analysis.

C0788: **Bayesian variable selection for skew normal models**
Presenter: **Andriette Bekker**, University of Pretoria, South Africa
Co-authors: Mohammad Arashi, Janet Van Niekerk, Arnold van Wyk

Variable selection is one of the most commonly faced problems in statistical analysis. In the frequentist paradigm, penalized regression methods such as L1 regularization and LASSO are used to induce sparsity in high-dimensional settings. In the Bayesian setting, sparsity can be induced by means of a two-component mixture prior with sufficient probability mass at zero. There has also been a recent development that uses global-local shrinkage priors for high-dimensional Bayesian variable selection. The Dirichlet-Laplace (DL) prior is a popular example of this and has shown promising results compared to existing feature selection methods in the Bayesian framework. Incorporating an asymmetrical component into the variable selection framework is proposed. This is showcased by incorporating a skew-normal random error component into the Dirichlet-Laplace prior to linear regression. A framework for prior selection and hyperparameter tuning of the proposed model is also proposed. The performance of the proposed model is assessed and compared with its symmetrical counterpart in both simulated and real-data examples. It is found to not only perform well but also identify certain non-zero signals due to the inclusion of skewness in the proposed model.

C1279: **Evaluation of the random match probability in forensic statistics**
Presenter: **Gianfranco Piscopo**, University of Naples Federico II, Italy
Co-authors: Maria Longobardi, Massimiliano Giacalone

Forensic statistics is generally referred to as the detection of data on crimes and trial outcomes, with subsequent analysis of the data thus detected. Statistics applied to the evaluation of evidence is increasingly being recognized as an important part of the modern criminal justice system. The use of DNA tests for identification and judicial purposes has been the greatest revolution in criminal investigation. It is founded on the Bayesian causal inference and on the evaluation of the random match probability. Forensic science is the application of scientific methodologies to traditional investigative techniques in relation to the detection of a crime. The applicative field is wide, ranging from biology to psychology, chemistry, computer science, physics, medicine and engineering. Forensic genetics is a specialized branch of forensic science that analyzes DNA to prove suspicious responsibility for a crime. The aim is to consider the projection pursuit approach in forensic genetics by providing new tools and techniques that help deliver justice in an increasingly complex world. This is made possible by the high dimensionality of the data, which allows the use and comparison of techniques developed in very different inferential contexts.

CO043 Room BH (SE) 2.09 APPLIED MACRO-FINANCE I Chair: Alessia Paccagnini

C1364: **Paying for the prices: The cost of taming inflation**
Presenter: **Andre Casalis**, National Bank of Slovakia, Slovakia

Using high-frequency data on individual bank account transactions and card payments, the impact of monetary policy is investigated on consumption at daily frequency, and the focus is on the magnitude and transmission dynamics of interest rate shocks. The granularity of the data allows building consumption series segmented by age, gender, education level, and region to explore asymmetric features of monetary policy transmission. A parsimonious local projection specification allows flexible inclusion of a variety of controls to explore the consumption effects of the full maturity profile of the yield curve, and to disentangle extensive and intensive margins contribution. Furthermore, a nonlinear extension of the framework is able to identify the effects of positive and negative monetary policy shocks. A selection of findings includes: (a) household spending reaction peaks approximately eleven months after the initial shock; (b) negative shocks are definitely contractionary, while positive shocks are unable to show a decisive expansionary effect; (c) interest rate shocks from the short and medium-term maturities do not present significant differences in the way in which they affect consumption, while longer term maturities transmits quicker to household spending; (d) monetary policy is symmetric in its effects and transmission timing across the demographic dimensions of age, sex, education, and region.

C1379: **Club convergence of real wages in the European Union**
Presenter: **Vladimir Arcabic**, Faculty of Economics and Business, Croatia

Co-authors: Tomislav Globan, Goran Markusic

Real wage convergence is analyzed within the European Union, scrutinizing it on both regional and national scales. It questions the prevalent notion that open labor movement within the EU naturally leads to wage parity. Using a flexible log t regression and clustering algorithm, the research uncovers a lack of absolute wage convergence across EU countries and regions. Instead, it identifies distinct convergence clubs based on real wages, delineated as core versus peripheral countries, and further divided into four regional categories. These findings suggest the existence of significant barriers, like cultural and linguistic differences, impeding the free movement of labor and thereby hindering wage equalization. Clubs converge to different steady states, which yield substantially different average real wages. At the national level, the formation of these wage-based clubs is influenced by the output gap, labor market regulations, and productivity. The results are robust and consistent with economic theories concerning real wage determinants. On the regional front, the analysis reveals a complex interplay of multiple variables affecting club formation, with no significant distinctions between certain club pairings. This research challenges established economic theories and underscores the intricate relationship between economic policies, cultural factors, and wage convergence in the European Union.

C1478: **Climate risks and sovereign risks nexus**

Presenter: **Marianna Blix Grimaldi**, Swedish Riksbank, Sweden

Co-authors: Sofia Anyfantaki, Carlos Madeira, Simona Malovana, Georgios Papadopoulos

Governments are exposed to a significant range of climate-related risks, which can, in turn, affect government spending and the value of government bonds. Surprisingly, few studies have thoroughly analyzed the effects of these climate risks on sovereign debt. The impact of both physical and transition risks on sovereign bond yields is investigated, making a clear distinction between chronic and acute physical risks. By employing panel regression models on a comprehensive dataset of 57 countries, the influence of various climate risks is estimated, and the local projection method is used to examine physical risks in greater detail. Findings reveal that both transition risks and chronic physical risks generally lead to an increase in government bond yields, although the extent of this impact varies considerably across different countries and regions. Interestingly, severe, short-term physical risks appear to have minimal influence on sovereign bond yields in emerging and developing economies. This suggests that financial markets may not yet fully account for climate risks in their pricing. The importance of investigating sovereign bond markets is underscored by the fact that these bonds often constitute a significant portion of asset portfolios. As such, understanding the intersection between climate risks and sovereign debt is crucial for investors and policymakers alike.

C0211: **Oil shocks and firm-level expectations**

Presenter: **Petre Caraiani**, Bucharest University of Economic Studies, Romania

Employing firm-level data from Germany, the aim is to investigate the influence of oil news on firms' economic expectations, specifically focusing on price and employment expectations. Through panel data analysis, it is demonstrated that oil news shocks precipitate a significant increase in firms' price expectations while concurrently diminishing their employment projections. Notably, the peak effect on price expectations manifests within 3-4 quarters, whereas employment expectations adjust more rapidly within just one month. The robustness of these findings is found across various types of oil news shocks and different firm sub-samples, underscoring the pervasive impact of oil news on economic forecasts at the firm level. The results also indicate new evidence for expectation formation, as the peak responses are reached after 3-4 months for prices and 6-9 months for employment. Oil news shocks also have a more lasting impact on employment expectations.

CO025 **Room BH (SE) 2.12** SPATIAL STATISTICS: COMPUTER EMULATION, NEUROSCIENCE & ECOLOGY **Chair: Rajarshi Guhaniyogi**

C1171: **A bootstrap-based goodness of fit test for binary spatial models**

Presenter: **Eva Biswas**, Iowa State University, United States

Co-authors: Daniel Nordman, Mark Kaiser, Andee Kaplan

Binary spatial observations frequently occur in environmental and ecological studies, where Markov random field (MRF) models are commonly applied. Despite their widespread use and long-standing history, appropriate model diagnostics for spatial binary data in MRF models have remained a challenging issue. A complicating factor is the difficulty in assessing neighborhood specifications for binary data. To address this, a formal goodness-of-fit test is proposed for diagnosing MRF models for spatial binary values. The test statistic involves a type of conditional Moran's I based on the fitted conditional probabilities, which is capable of detecting departures in model form, including neighborhood structure. The application of the spatial test is illustrated using a dataset on the breeding pattern of grasshopper sparrows across Iowa.

C1175: **Unraveling complex relationships between misaligned images with additive neural network Gaussian processes**

Presenter: **Rene Gutierrez**, University of Texas at El Paso, United States

Co-authors: Rajarshi Guhaniyogi, Aaron Scheffler

The presentation explores relationships among images of varying scales, resolutions, and shapes, addressing a key issue in image-based data. The focus is on misalignment from data captured at different scales: upper-level images segmented into regions and lower-level images segmented into sub-regions. Within a regression framework, the response region and a subset of predictor images are defined at the lower scale, with an additional predictor image derived from a network where nodes represent upper-level regions. A regression framework that captures non-linear effects of network and lower-level predictors on responses is introduced, using a Gaussian process (GP) prior with a neural network-based covariance function (NN-GP prior). This framework merges Bayesian GP regression's uncertainty quantification with neural networks' predictive power, offering a flexible non-linear regression approach. The method enables scalable computation, captures intricate characteristics often missed by local spatial smoothing methods, and improves robustness by treating lower-scale image sub-regions as effective samples. Simulation studies show that the approach outperforms existing methods in predictive inference for response images.

C1321: **Exact MCMC-free Bayesian inference for data of any size**

Presenter: **Jonathan Bradley**, Florida State University, United States

Fine particulate matter and aerosol optical thickness are two variables of interest to scientists for understanding air quality and its various health and ecological impacts. Data on these variables are extremely large, making Bayesian analysis impractical. Scalable exact posterior regression (S-EPR) is introduced, which combines two recently introduced methodologies: the data subset approach and exact posterior regression (EPR). The "data subset approach" assumes a parametric model for a low-dimensional training dataset and assumes the remaining holdout data follows its true data-generating mechanism. Posterior samples from this model scale to the low-dimensional training data while simultaneously including all the available data, making Bayesian inference from this model scalable to arbitrary dimensions. The data subset approach is combined with a Bayesian hierarchical model that allows one to sample independent replicates of fixed and random effects directly from the posterior without the use of MCMC or approximations. Samples from the posterior distribution have an efficient projection form and, hence, are referred to as EPR. For the first time, an exact, fully Bayesian method for a class of spatial GLMMs can be scaled to arbitrary dimensions and does not require MCMC. The benefits of S-EPR are illustrated via the motivating application.

C1640: **Comparing emulators systematically**

Presenter: **Devin Francom**, Los Alamos National Laboratory, United States

Many emulation methods exist, and no one emulator works best for all situations. A systematic comparison of a collection of emulators is provided. The accuracy of predictions and uncertainty estimates is compared using a broad set of test functions and computer model datasets (all with scalar response). The comparison includes emulation methods like Gaussian processes, Bayesian additive regression trees, Bayesian multivariate adaptive regression splines, Bayesian projection pursuit regression, Bayesian neural networks, and Bayesian polynomial chaos. Further, an R package, duqling, is introduced to make the comparison reproducible. For example, if someone wonders how their favorite emulation method would compare (or one of the above methods under different settings), they can train and test under the exact same conditions of this analysis by using the R package.

CC507 Room BH (SE) 2.05 COMPUTATIONAL AND FINANCIAL ECONOMETRICS **Chair: Antoine Djogbenou**

C1717: **High-dimensional covariance matrix estimators on simulated portfolios**
Presenter: **Andres Garcia**, Autonomous University of Baja California, Mexico

The allocation of synthetic portfolios under different dependency structures is studied in a high-dimensional context. The research tests approaches based on random matrices, free probability, deterministic equivalents, and their combination with hierarchical clustering. Simulations are compared with the out-of-sample performance of empirical data from the companies that make up the S&P 500 index, evaluating metrics such as annual return, annual volatility, Sharpe ratio, maximum drawdown, Sortino ratio, and turnover. The portfolio allocation strategies analyzed include the minimum variance portfolio, both with and without short-selling constraints, as well as the hierarchical risk parity approach. The results pave the way for new risk management proposals.

C1716: **Identification and estimation of network models using panel data analysis**
Presenter: **Vasilis Sarafidis**, Brunel University London, United Kingdom
Co-authors: George Kapetanios

A novel methodology is proposed for identifying and estimating spatial and network models using large panel data. We address the challenge of estimating the spatial interactions between individual units by developing a boosting algorithm that relies on the statistical significance of individual neighbouring covariates tested one at a time. We refer to the proposed method as Boosting One Neighbour at a Time Multiple Testing (BONMT) procedure. Our approach allows for the flexible selection of neighboring units in the presence of high-dimensional networks, even in cases where the cross-sectional dimension of the panel is larger than the number of time series observations available. Furthermore, our procedure is robust to unequal weights, i.e., asymmetric network systems where some individuals influence their neighbors more strongly than others. Theoretical results are provided, demonstrating the algorithm's consistency and asymptotic properties under various network structures. The resulting IV estimator is shown to be consistent as N,T both grow large. Monte Carlo simulations illustrate the methods excellent finite-sample performance, confirm its robustness across different network configurations and levels of spatial correlation and show that it can outperform alternative methods.

C1713: **Comparative evaluation of open-source and proprietary software for ARDL, EC models, and bounds test for cointegration**
Presenter: **Kleanthis Natsiopoulos**, University of Dundee, United Kingdom
Co-authors: Nickolaos Tzeremes

The Autoregressive Distributed Lag (ARDL) and Error Correction (EC) models are widely used in time-series econometrics, particularly for analyzing level (cointegrating) relationships between variables in the presence of mixed levels of integration. Since its release, the ARDL package for R has become a reliable and versatile tool for implementing these models, offering an intuitive framework for estimating ARDL and EC models, automatically selecting optimal lags and conducting bounds tests for cointegration. The package also provides seamless integration with other R packages for post-estimation diagnostics, further enhancing its utility for applied econometric analysis. Additionally, a proof-of-concept (PoC) comparison of the ARDL package with other open-source R packages and proprietary software (EViews, Stata) is presented. The ARDL package demonstrates accurate and consistent results, surpassing other open-source alternatives and offering several advantages over proprietary tools, such as exact sample critical values and additional model representations. The comparison uses the original data from the seminal bounds test paper.

C1722: **Multidimensional arrays, indices and Kronecker products**
Presenter: **Stephen Pollock**, University of Leicester, United Kingdom

Much of the algebra that is associated with the Kronecker product of matrices has been rendered in the conventional notation of matrix algebra, which conceals the essential structures of the objects of the analysis. This makes it difficult to establish even the most salient of the results. The problems can be greatly alleviated by adopting an orderly index notation that reveals these structures. This claim is demonstrated by considering a problem that several authors have already addressed without producing a widely accepted solution.

CC474 Room BH (SE) 2.10 MACHINE LEARNING FOR ECONOMICS AND FINANCE I **Chair: Ines M del Puerto**

C0774: **Sparse neural networks and explainability in financial statement analysis**
Presenter: **Lars Fluri**, University of Basel, Switzerland

An alternative approach is proposed to feature selection and sparse modelling in the context of financial data analysis to predict free cash flow. Utilising deep learning important features (DeepLIFT), a process is introduced for iterative elimination of input features and extends currently used algorithms. This reduces the model complexity and enhances the robustness through the elimination of less significant input nodes. Furthermore, a method for regrowth of nodes using gradient magnitude of previously eliminated features based on state-of-the-art methods is used. Drawing on a dataset of 874 firms from the DACH region over a decade, the model is used to identify forward-looking predictors of free cash flow. Additionally, it evaluates both computational aspects and performance metrics (including in-sample and out-of-sample performance) to measure improvements from the original dense model to the optimized sparse model. By reducing the number of total features present by over 75% (from 49 to 10 features), out-of-sample R^2 decreases by only 13% while the standard deviation is improved by over 35%. The contribution is to the evolving field of machine learning applications in finance, proposing an alternative framework for feature selection and model optimization.

C1310: **A dynamic assessment of fairness on open banking data**
Presenter: **Yujia Chen**, University of Edinburgh Business School, United Kingdom
Co-authors: Raffaella Calabrese

The detailed nature of transaction data shared through open banking poses risks, as subtle proxies for sensitive and prohibited characteristics may lead to indirect discrimination. A novel dynamic methodology is introduced for assessing fairness in a loan approval process that takes the time of the transaction into account. This approach also identifies the variables responsible for any potential bias when utilizing dynamic transaction data alongside advanced technologies, such as machine learning.

C1673: **Smoothness of directed chain stochastic differential equations and its applications**
Presenter: **Tomoyuki Ichiba**, University of California Santa Barbara, United States

On a filtered probability space for the space of continuous functions, a system of stochastic equations shall be considered: directed chain stochastic differential equations for a pair of stochastic processes whose marginal distributions in the path space are identical and their joint distribution is uniquely determined by the system of equations with the distributional constraints. The smoothness of the solutions is discussed under some regular conditions first, and some relaxation of the conditions is then considered on the coefficients and the distributional constraints. Its applications are also introduced in such systems in the stochastic filtering problem and in the generative adversarial network problem.

C1387: **Reinforcement learning for credit risk**
Presenter: **Jorge C-Rella**, University of A Coruna, Spain

Co-authors: Juan Vilar Fernandez, Ricardo Cao, David Martinez Rego

The problem of credit risk is dynamic, as customers' payment behavior evolves according to the economic cycle and market trends, and cost-sensitive, as the results depend on the amount of the loan. In addition, information is only available for approved transactions, which can lead to unfair biases and opportunity costs. New dynamic learning strategies are proposed that extend online learning and bandit algorithms to cost-sensitive learning. Experiments on simulations and real-world datasets demonstrate the effectiveness of the proposed algorithms, opening the door to extending new methods to credit risk and other cost-sensitive problems.

Sunday 15.12.2024 17:40 - 18:55 Parallel Session J – CFECMStatistics2024

CO138 Room S-2.23 TRUSTWORTHY AI (VIRTUAL) Chair: Wei Sun

C0251: **A plug-and-play watermark framework for AI-generated images**
Presenter: **Xuan Bi**, University of Minnesota, United States

Safeguarding intellectual property and preventing potential misuse of AI-generated images are of paramount importance. A robust and agile plug-and-play watermark detection framework, dubbed RAW, is introduced. As a departure from traditional encoder-decoder methods, which incorporate fixed binary codes as watermarks within latent representations, the approach introduces learnable watermarks directly into the original image data. Subsequently, a classifier is employed that is jointly trained with the watermark to detect the presence of the watermark. The proposed framework is compatible with various generative architectures and supports on-the-fly watermark injection after training. By incorporating state-of-the-art smoothing techniques, the framework provides provable guarantees regarding the false positive rate for misclassifying a watermarked image, even in the presence of certain adversarial attacks targeting watermark removal. Experiments on a diverse range of images generated by state-of-the-art diffusion models reveal substantial performance enhancements compared to existing approaches. For instance, the method demonstrates a notable increase in AUROC, from 0.48 to 0.82, when compared to state-of-the-art approaches in detecting watermarked images under adversarial attacks while maintaining image quality, as indicated by closely aligned FID and CLIP scores.

C0905: **Synergetic random forests for policy evaluation and uncertainty quantification in reinforcement learning**
Presenter: **Ruoqing Zhu**, University of Illinois at Urbana-Champaign, United States
Co-authors: Zexuan Zhang, Rui Qiu, Zhou Yu

A new breed of random forest model is proposed, in which the splitting rule depends not only on the within-node data but also on information across the entire tree. The new model is particularly suited for situations when the splitting rule, while viewed as an estimating equation, requires further estimation of nuisance parameters that are not feasible within the node. In the proposed model, the nuisance parameter estimation is synergized across the entire tree and also progressively grows as tree nodes expand, facilitating the estimation of the main parameter of interest. A typical use case of such a model is policy evaluation in reinforcement learning when estimating the value function can utilize information from the transitional kernel. Utilizing the platform of random forests, the uncertainty of policy evaluation can also be easily quantified, which can often be challenging with other approaches.

C1070: **Sequential conformal prediction for time series**
Presenter: **Chen Xu**, Georgia Institute of Technology, United States

A new distribution-free conformal prediction algorithm is presented for sequential data (e.g., time series), called the sequential predictive conformal inference (SPCI). The fact that time series data are non-exchangeable is specifically accounted for, and thus, many existing conformal prediction algorithms are not applicable. The main idea is to adaptively re-estimate the conditional quantile of non-conformity scores (e.g., prediction residuals) upon exploiting the temporal dependence among them. More precisely, the problem of conformal prediction interval is cast as predicting the quantile of a future residual, given a user-specified point prediction algorithm. Theoretically, asymptotic valid conditional coverage is established upon extending consistency analyses in quantile regression. Using simulation and real-data experiments, a significant reduction is demonstrated in the interval width of SPCI compared to other existing methods under the desired empirical coverage. Extensions to multivariate time series are also discussed.

CO382 Room S-2.25 RECENT APPROACHES IN SPECIES DISTRIBUTION MODELLING Chair: Maria Franco Villoria

C0768: **Bayesian inference on high-dimensional multivariate binary responses**
Presenter: **Antik Chakraborty**, Purdue University, United States

It has become increasingly common to collect high-dimensional binary response data, for example, with the emergence of new sampling techniques in ecology. In smaller dimensions, multivariate probit (MVP) models are routinely used for inferences. However, algorithms for fitting such models face issues in scaling up to high dimensions due to the intractability of the likelihood, involving an integral over a multivariate normal distribution having no analytic form. Although a variety of algorithms have been proposed to approximate this intractable integral, these approaches are difficult to implement and/or inaccurate in high dimensions. The focus is on accommodating high-dimensional binary response data with a small-to-moderate number of covariates. A two-stage approach is proposed for inference on model parameters while taking care of uncertainty propagation between the stages. The special structure of latent Gaussian models is used to reduce the highly expensive computation involved in joint parameter estimation to focus inference on marginal distributions of model parameters. This essentially makes the method embarrassingly parallel for both stages. Performance in simulations and applications is illustrated in joint species distribution modeling in ecology.

C1010: **Variance partitioning-based priors for species distribution models**
Presenter: **Luisa Ferrari**, University of Bologna, Italy
Co-authors: Massimo Ventrucci

Species distribution models for community ecology often require complex models to capture the influence of numerous abiotic factors with complex relationships, as well as spatial and temporal patterns in the data. For this reason, the Bayesian hierarchical framework is a popular choice for analysing species occurrence data. However, traditional prior specification approaches have been found to have several disadvantages. Recently proposed variance partitioning-based priors offer a promising new framework to handle the problem of incorporating prior assumptions, particularly for complex models. The core idea of this prior class consists of designing a more intuitive reparametrization of the original parameters for which specification becomes easier for the user. The applicability is, however, so far limited to models that include a subset of all possible effects. The purpose is to explore how to extend this new method to a wider class of models. In particular, it focuses on incorporating spatial and temporal smoothing effects, which are highly popular within species distribution models. The aim is to expand the usefulness of variance-partitioning-based priors to new fields of application, such as the context of ecology studies, among others.

C1325: **Auto-correlation-driven environmental sampling: an adaptive approach**
Presenter: **Linda Altieri**, University of Bologna, Italy
Co-authors: Daniela Cocchi

Spatially balanced techniques are commonly used in environmental sampling. These methods return well-spread samples across the study area, which is considered a highly efficient approach for estimating the total or mean of a variable of interest. We have shown that such techniques perform well only if the variable is positively correlated over space. Conversely, in environmental data, it is common to encounter negative correlations, such as when plants or animals compete for soil or natural resources. We propose a new approach to spatial sampling, which sequentially adapts the selection of units to the type of correlation of the study variable. Our method is based on Spatially Correlated Poisson Sampling, incorporating a novel weighting system that either encourages or discourages the selection of units at specific distances, depending on the strength and nature of the correlation at those distances. The estimation of this association is adaptively refined as the sampling progresses. When the spatial arrangement follows a random or negatively correlated structure, our approach results in estimates with a significantly smaller estimation error. We demonstrate the effectiveness of this method through a case study using ecological data.

CO351 Room S-1.01 HiTEc: Model instabilities and data dependency Chair: Matus Maciak

C0871: **Exogenous and endogenous market effects: Model based change-point detection**
Presenter: **Matus Maciak**, Charles University, Czech Republic

Various options market turmoil are analyzed over time using artificial interpolated volatilities carefully constructed in order to distinguish between exogenous and endogenous market effects. A panel data quantile regression model is further applied to detect significant change points occurring in the options market behavior. Some necessary statistical theories are discussed together with computational and algorithmic details. Finite sample performance is assessed using empirical comparisons, together with a real data illustration.

C1394: **Changing intensities**
Presenter: **Michal Pesta**, Charles University, Czech Republic
Co-authors: Marie Huskova

The focus is on situations such that some phenomena can cause several related events, and each event contributes to a different univariate counting process. A collection of these naturally dependent point processes, therefore, forms a flexible multivariate counting process, where neither stationarity nor independence of interarrival times of the marginal processes are assumed. The main aim is to detect a structural break of the event's occurrences over time, which means to test whether some (not necessarily all) intensities of the univariate counting processes are subject to change at some unknown time point.

C1399: **Changepoint detection in tensor data**
Presenter: **Barbora Pestova**, Charles University, Czech Republic
Co-authors: Michal Pesta, Martin Romanak

Tensor data consisting of multivariate outcomes over the items and across the subjects with longitudinal and cross-sectional dependence are considered. Distributional-free detection procedures for changepoints at different locations are proposed, which are in an unsupervised learning manner. The bootstrap superstructure is developed to overcome computational issues in such a universal setup. The completely data-driven test is presented using real data examples.

CO508 Room S-1.04 Advances in data depth Chair: Stanislav Nagy

C1723: **Halfspace depth for directional data**
Presenter: **Stanislav Nagy**, Charles University, Czech Republic
Co-authors: Rainer Dyckerhoff

The angular halfspace depth (ahD) is a natural modification of the celebrated halfspace (or Tukey) depth to the setup of directional data. It allows us to define elements of nonparametric inference, such as the median, the inter-quantile regions, or the rank statistics, for datasets supported in the unit sphere. Despite being introduced previously, ahD has never received ample recognition in the literature, mainly due to the lack of efficient algorithms for its computation. We address both the computation and the theory of ahD. First, we express the angular depth ahD in the unit sphere in the d-dimensional space as a generalized (Euclidean) halfspace depth in dimension d-1, using a projection approach. That allows us to develop fast exact computational algorithms for ahD in any dimension d. Second, we show that similarly to the classical halfspace depth for multivariate data, also ahD satisfies many desirable properties of a statistical depth function. Further, we derive uniform continuity/consistency results for the associated set of directional medians, and the central regions of ahD, the latter representing a depth-based analog of the quantiles for directional data.

C1725: **Explicit bivariate simplicial depth**
Presenter: **Erik Mendros**, Charles University, Czech Republic
Co-authors: Stanislav Nagy

The simplicial depth (SD) is a celebrated tool defining elements of nonparametric and robust statistics for multivariate data. While many properties of SD are well-established, and its applications are abundant, explicit expressions for SD are known only for a handful of the simplest multivariate probability distributions. In our presentation, we deal with SD in the plane. We start by developing a one-dimensional integral formula for SD of any properly continuous probability distribution. We then apply this formula to derive exact, explicit expressions for SD of uniform distributions on (both convex and non-convex) polygons in the plane or on the boundaries of such polygons. Additionally, we discuss several implications of these findings to probability and statistics: (a) An upper bound on the maximum SD in the plane, (b) an implication for a test of symmetry of a bivariate distribution, and (c) a connection of SD with the classical Sylvester problem from geometric probability.

C1724: **Halfspace depth for alpha-symmetric distributions**
Presenter: **Filip Bocinec**, Charles University, Czech Republic
Co-authors: Stanislav Nagy

The well-studied halfspace depth is a valuable nonparametric tool for assessing the centrality of points in relation to multivariate probability distributions. Extending this concept, scatter halfspace depth applies to scatter matrices, enabling the analysis of variability. The convergence rates of both location and scatter halfspace medians for contaminated elliptical distributions have been previously analyzed, demonstrating that these medians achieve minimax optimality in such models. We investigate properties of both location and scatter halfspace depths for a broader class of distributions, known as alpha-symmetric distributions, with a focus on the convergence rates of their medians. Extending some previous results, we address the general family of alpha-symmetric distributions, encompassing both elliptically symmetric (alpha = 2) and multivariate heavy-tailed (alpha < 2) cases. We establish a convergence rate for the location halfspace median of an alpha-symmetric distribution, while showing that an analogous result for the scatter halfspace median is feasible only under elliptical symmetry (alpha = 2).

CO337 Room S-1.06 Interpretable machine learning and high-dimensional statistics (virtual) Chair: Garvesh Raskutti

C0293: **Cluster quilting: Spectral clustering for patchwork learning**
Presenter: **Lili Zheng**, University of Illinois Urbana-Champaign, United States
Co-authors: Andersen Chang, Genevera Allen

Patchwork learning arises as a new and challenging data collection paradigm where both samples and features are observed in fragmented subsets. Due to technological limits, measurement expense, or multimodal data integration, such patchwork data structures are frequently seen in neuroscience, healthcare, and genomics, among others. Instead of analyzing each data patch separately, it is highly desirable to extract comprehensive knowledge from the whole data set. The focus is on the clustering problem in patchwork learning, aiming at discovering clusters amongst all samples even when some are never jointly observed for any feature. A novel spectral clustering method is proposed called cluster quilting, consisting of (i) Patch ordering that exploits the overlapping structure amongst all patches, (ii) Patchwise SVD, (iii) Sequential linear mapping of top singular vectors for patch overlaps, followed by (iv) K-means on the combined and weighted singular vectors. Under a sub-Gaussian mixture model, theoretical guarantees are established via a non-asymptotic misclustering rate bound that reflects both properties of the patch-wise observation regime as well as the clustering signal and noise dependencies. The cluster quilting algorithm is also validated through extensive empirical studies on both simulated and real data sets in neuroscience and genomics, where it discovers more accurate and scientifically more plausible clusters than other approaches.

C1189: **Can large language models solve compositional tasks: A study of out-of-distribution generalization**
Presenter: **Yiqiao Zhong**, UW Madison, United States

Large language models (LLMs) such as GPT-4 sometimes appeared to be creative, solving novel tasks with a few demonstrations in the prompt. These tasks require the pre-trained models to generalize on distributions different from those from training data - which is known as out-of-distribution generalization. For example, in "symbolized language reasoning", where names/labels are replaced by arbitrary symbols, the model can infer the names/labels without any fine-tuning. The focus is on a pervasive structure within LLMs known as induction heads. By experimenting on a variety of LLMs, it is empirically demonstrated that compositional structure is crucial for Transformers to learn the rules behind training instances and generalize on OOD data. Further, the "common bridge representation hypothesis" is proposed, where a key intermediate subspace in the embedding space connects components of early layers and those of later layers as a mechanism of composition.

C1204: **Optimal iterative algorithms for structured PCA with invariant noise**
Presenter: **Rishabh Dudeja**, University of Wisconsin–Madison, United States
Co-authors: Songbin Liu, Junjie Ma

The problem of recovering a low-rank signal matrix is considered from a noisy observed matrix corrupted with additive noise. When the noise matrix is i.i.d. Gaussian, a rich line of work has characterized the information-theoretic limits for this problem and determined the smallest possible estimation error achievable by computationally efficient estimators. The i.i.d. noise model constrains the eigenvalue spectrum of the observed matrix to follow the semi-circle law, which may not accurately represent all datasets. A flexible generalization of the i.i.d. Gaussian noise model, known as the rotationally invariant noise model, is studied, which can capture noise spectrums beyond the semi-circle law. A new class of approximate message-passing algorithms is developed for this problem, and their dynamics are characterized. These algorithms leverage prior knowledge about the noise and signal structures by iteratively applying non-linear denoisers to the eigenvalues of the observed matrix and the previous iterates. The optimal choices are identified for these denoisers, and evidence is provided, suggesting that the resulting algorithm is a natural candidate for the optimal computationally efficient algorithm by showing that it achieves the smallest possible estimation error among a broad class of iterative algorithms under a given iteration budget.

CO262 Room S-1.27 STUDY DESIGN AND CAUSAL INFERENCE ISSUES IN COMPLEX BIOMEDICAL STUDIES Chair: Florin Vaida

C1683: **Small-sample behavior of the test for comparing standardized mean differences in meta-analysis**
Presenter: **Anya Umlauf**, University of California San Diego, United States

Two common estimators for the difference between standardized means are Cohen's d and Hedges' g. For either estimate, a comparison of two effect sizes can be done with a test based on normal distribution assumption. Much has been written on test performance for large sample sizes (more than 10 per group). Group sizes below ten are not unusual, however, particularly in pilot studies or studies in areas with limited resources. Having a reliable test for such a situation would be desirable. The test for comparing two effect sizes is dependent on the variances for the effect sizes being compared. There are several ways of estimating the effect size variance for both Cohen's d and Hedges' g. We set out to investigate which method leads to the most accurate results when group sizes are small. We simulated tests under normality assumption using gamma-based estimates, as well as the large-sample approximation and unbiased estimate, both proposed by Hedges. We also tested a pooled estimate calculated from the unbiased variance estimates. We expected the approximated statistics to yield biased results for small samples, but the simulations showed the large-sample estimate outperformed other approximations. Tests based on Hedges' g were more accurate than the tests for Cohen's d.

C1694: **Estimators of variance for matching-based estimates in the setting of complex surveys**
Presenter: **Karen Messer**, University of California, San Diego, United States

The large sample properties of resampling estimates of variance for matching estimators of the ATT are studied in the setting of a large-scale complex population survey. Resampling-based estimators of variance must reflect the complex hierarchical sampling design of the survey. Our working example is the Current Population Survey from the US Census Bureau, which uses published replicate weights to implement a jackknife estimate of variance, using the computational refinement of Balanced Repeated Replication, a common approach. Other approaches to these data include bootstrap-based methods, either conditional on the match or incorporating the randomness of the matching process. We give an overview of methodological issues in this setting, and present some recent work.

C1693: **Design issues for longitudinal, cluster randomized clinical trials with repeated measures**
Presenter: **Florin Vaida**, University of California San Diego, United States

The aim is to investigate design issues for cluster-randomized, longitudinal clinical trials with repeated measures. Units grouped within clusters are randomized to two or more groups. Units are observed longitudinally, at baseline and follow-up visits. Repeated measures for the response of interest are obtained at each visit. We show that a specific imbalance allocation of measures between baseline and follow-up is optimal. The robustness of the design is also considered. Statistical analyses are based on analytical derivations. We show that the optimal analyses are those controlling for baseline as a covariate, or including baseline in a longitudinal analysis and assuming no baseline differences. We show analytically that the two approaches are equivalent for finite samples. This approach was applied to the Open & Ask Study (NCT ID NCT03385512), a recently published US large-scale multi-center cluster-randomized controlled trial assessing the comparative effectiveness of three interventions to improve engagement of healthcare providers with their patients.

CO032 Room Auditorium DYNAMIC ECONOMETRICS Chair: Esther Ruiz

C0178: **Dealing with idiosyncratic cross-correlation when constructing confidence regions for PC factors**
Presenter: **Esther Ruiz**, Universidad Carlos III de Madrid, Spain
Co-authors: Diego Fresoli, Pilar Poncela

The finite sample performance of asymptotic and bootstrap regions is analysed for PC factors. It is shown that when the idiosyncratic components are wrongly assumed to be cross-sectionally uncorrelated, prediction regions for the estimated factors based on standard asymptotic results can have wrong coverages, which can be either larger or smaller than the nominal depending on the covariances of the idiosyncratic noises and the factor loadings. Procedures to compute the asymptotic MSE of the factors, taking into account the idiosyncratic cross-dependence, can help but are still inadequate depending on the structure of the cross-correlations. It is also shown that alternative extant bootstrap procedures may also have wrong coverages in front of realistic idiosyncratic correlations. Alternatively, a computationally simple estimator of the asymptotic covariance matrix of the factors is proposed based on adaptive thresholding of the sample covariances of the idiosyncratic residuals with the threshold based on the variance of each individual entry of the sample covariances.

C0198: **A new approach to regime switching autoregressions**
Presenter: **Frederik Krabbe**, Aarhus University, Denmark
Co-authors: Leopoldo Catania, Andrew Harvey

A new way is discussed to construct regime-switching autoregressions, making use of a non-Markovian unobserved process. It is shown that, in a special case, the likelihood implied by this new specification is identical to the classical Markov switching autoregression one. The general case leads to more flexible specifications with tractable likelihood functions. The statistical properties of the model are discussed, and conditions for the consistency and asymptotic normality of the maximum likelihood estimator are established. An application to macroeconomic variables shows that the new specification leads to better estimates and predictions.

C0479: **Skewness and kurtosis of aggregated financial returns**
Presenter: **Angeles Carnero**, Universidad de Alicante, Spain
Co-authors: Angel Leon, Trino Niguez

Analytical expressions are provided for the mean, variance, skewness and kurtosis of non-overlapping aggregated returns generated by a TGARCH(1,1) model assuming different skewed distributions for the innovations, including those represented by polynomially adjusted densities. Closed-form expressions for the skewness of aggregated returns present an alternative approach to approximating unconditional skewness. This contrasts with alternative methods employed in the literature, which rely on a second-order Taylor series expansion within the context of GARCH-type models. Aggregated moments facilitate the derivation of predictions for multiple-period value-at-risk and expected shortfall, and they can also be used in option pricing models.

CO286 Room K0.16 NETWORK SCIENCE IN PUBLIC HEALTH Chair: Thien Minh Le

C0742: **A generalized estimating equation approach to network regression**
Presenter: **Riddhi Pratim Ghosh**, Bowling Green State University, United States
Co-authors: Jukka-Pekka Onnela, Ian Barnett

Modeling the spread of infectious diseases, such as COVID-19, through a network of individuals, hospitals, or countries poses methodological challenges. As has been well studied, naive regression neither properly accounts for network community structure nor does it account for the dependent variable acting as both model outcome and covariate. To address this methodological gap, a proposed network regression model is motivated by the important observation that controlling for community structure can, when a network is modular, significantly account for a meaningful correlation between observations induced by network connections. A generalized estimating equation (GEE) approach is proposed to learn model parameters based on node clusters defined through any single-membership community detection algorithm applied to the observed network. A necessary condition is provided on the network size and edge formation probabilities to establish the asymptotic normality of the parameters under the stochastic block framework. The approach is used to estimate the impact of the commercial air transportation network between countries on the spread of COVID-19 incidence rates as well as on the receipt of aid between countries. It is found that while naive regression overstates the significance of network effects post-lockdown, our approach more accurately quantifies the impact of the travel network on COVID-19 incidence rates.

C0821: **Creating more equitable access to care through interhospital transfer networks**
Presenter: **Korilyn Zachrison**, Massachusetts General Hospital and Harvard Medical School, United States

There are a number of inequities in access to high-quality healthcare, in large part related to differences in where resources are located relative to where patients live and seek care. While one cannot move where hospitals are located, patients can be brought to a higher level of resources through interhospital transfer. The connections that are made between hospitals through patient transfer can be understood as a network. This network of interhospital transfers connects hospitals by moving patients to resources required for their care. The current state of interhospital transfer networks is not maximized for optimal patient benefit. However, by applying strategies from network science, such networks may be leveraged in a number of ways to create more efficient and more equitable systems of medical care.

C1311: **Connecting mass-action models and network models for infectious diseases**
Presenter: **Thien Minh Le**, University of Tennessee at Chattanooga, United States

Infectious disease modeling is used to forecast epidemics and assess the effectiveness of intervention strategies. Although the core assumption of mass-action models of homogeneously mixed population is often implausible, they are nevertheless routinely used in studying epidemics and provide useful insights. Network models can account for the heterogeneous mixing of populations, which is especially important for studying sexually transmitted diseases. Despite the abundance of research on mass-action and network models, the relationship between them is not well understood. The attempt is to bridge the gap by first identifying a spreading rule that results in an exact match between disease spreading on a fully connected network and the classic mass-action models. A method for mapping epidemic spread on arbitrary networks to a form similar to that of mass-action models is then proposed. A theoretical justification for the procedure is also provided. Finally, the advantages of the proposed methods are shown using synthetic data that is based on an empirical network. These findings help in understanding when mass-action models and network models are expected to provide similar results and identify reasons when they do not.

CO244 Room K0.18 ENVIRONMENTAL DATA INTEGRATION, ESTIMATION, AND MAPPING Chair: Sara Franceschi

C0426: **Air quality data fusion using fixed rank Kriging with estimates at municipal level**
Presenter: **Alessandro Fusta Moro**, University of Bergamo, Italy
Co-authors: Jacopo Rodeschini, Andrea Moricoli, Alessandro Fasso

Within the ongoing Italian project "Growing resilient, inclusive, and sustainable" (GRINS), the need for a harmonised dataset containing all relevant variables at the municipal level has emerged. Implementing statistical models on this harmonised dataset about social, economic, and environmental data will allow researchers to gain meaningful insights from the data, providing a useful data-driven framework for policymakers. However, merging hundreds of different variables with different spatial and temporal supports and resolutions represents an important challenge. The purpose is to show how data fusion and the change of support problems are addressed within the statistical framework using the fixed rank Kriging model on air quality data. Air quality data come in different ways: chemical transport models (CTMs), national air quality monitoring network and European satellites (e.g. Sentinel 5P). Each source has its strengths and weaknesses and different spatial and temporal resolutions. It shows how to obtain harmonized data at the municipal level, respecting the peculiarities of each source and capitalizing on their strengths while mitigating their weaknesses. The method used further quantifies uncertainties along with predictions and provides intra-municipal information (e.g. population exposure curve within the same municipality).

C1112: **Improving spatial maps with preferential sampling via hierarchical modeling**
Presenter: **Giacomo Zoppi**, University of Torino, Italy
Co-authors: Natalia Golini, Rosaria Ignaccolo, Anna Lo Presti, Michela Cameletti

Mapping is essential to understanding the spatial pattern of species over a region. Recently, spatial modelling has been developed to explain the presence/absence or abundance of one or more species over a region, as well as through the use of environmental variables available at locations across the region. Nevertheless, the sampling location choice may not be completely random but guided by some kind of relationship with the variable of interest (preferential sampling). Misusing preferential sampling can introduce bias in the parameter estimates and in the predictions reported in the map. A hierarchical modelling approach is considered that takes into account preferential sampling for abundance data (e.g., counts, per cent cover, or biomass) and for the integration of abundance data with abundance-only data to improve prediction accuracy.

C1194: **Species coverage estimation by means of Monte Carlo integration techniques**
Presenter: **Agnese Marcelli**, University of Siena, Italy

Co-authors: Rosa Maria Di Biase, Sara Franceschi, Marzia Marcheselli, Caterina Pisani

Estimating species coverage is essential for understanding ecosystem health and effectively managing natural resources, requiring robust statistical techniques to ensure precise and reliable results. If the design-based approach is considered, the scheme for placing sample sites across a continuum is fundamental for performing reliable inference. In practical scenarios, such as estimating species coverage in dunes, strip sampling is a commonly adopted scheme, which involves recording the coverage of the target species within the strips. The aim is to show that, in this context, species coverage can be suitably expressed as an integral and, therefore, can be unbiasedly estimated using a Monte Carlo estimator. Additionally, a simulation study is presented to confirm the theoretical findings.

CO018 Room K0.19 LATEST TRENDS IN CLUSTERING AND CLASSIFICATION OF COMPLEX DATA II Chair: Marta Nai Ruscone

C1365: **Finite mixture of hidden Markov models for tensor variate time series data**
Presenter: **Xuwen Zhu**, The University of Alabama, United States
Co-authors: Abdullah Asilkalkan, Shuchismita Sarkar

The need to model data with higher dimensions, such as a tensor-variate framework where each observation is considered a three-dimensional object, increases due to rapid improvements in computational power and data storage capabilities. A finite mixture of hidden Markov models for tensor-variate time series data is developed. Simulation studies demonstrate high classification accuracy for both cluster and regime IDs. To further validate the usefulness of the proposed model, it is applied to real-life data with promising results.

C1463: **Finite mixture modeling for the analysis of spatiotemporal aspects in dendrochronology**
Presenter: **Volodymyr Melnykov**, The University of Alabama, United States
Co-authors: Lingge Wang

The objective of dendrochronology is to study environmental changes as well as date events and archaeological artifacts based on the sequences of annual rings extracted from tree trunks. Dendrochronology provides a unique yet scientifically sound approach to the analysis of events that occurred hundreds of years ago. The analysis of tree ring data is not trivial due to the presence of temporal relationships as well as the uncertainty associated with the location and lifetime of many trees. A finite mixture model is proposed to address the heterogeneity in the tree ring data. Unlike existing approaches discussed in the current literature, the proposed model takes into account both sources of uncertainty. The developed model is applied to the analysis of data publicly available from the international tree ring data bank, with promising results.

C1402: **Clustering time series of counts**
Presenter: **Luis Sousa**, University of Aveiro, Portugal
Co-authors: Isabel Pereira, Magda Monteiro

The clustering of time series has proven to be of interest in various fields, ranging from economics and finance to environment and medicine, among others. The objective is to group similar items according to a criterion suitable for the problem. Specifically, this aims to help outline strategies for better decision-making in the context of logistics, particularly in maritime ports. Some of the problems that may arise within this context range from predicting the number of ships arriving at one port to the clustering of ships based on the types of materials they are transporting. Much of the work developed over the recent decades has been conducted within the framework of continuous-valued time series, with few studies on clustering for count time series. The aim is to establish and apply model-based clustering to appropriately define discrete-valued time series, particularly those that allow for overdispersion and/or zero inflation. The idea is to use a finite mixture model that accommodates the mentioned characteristics, and several existing techniques, such as the selection of the number of clusters, estimation using expectation-maximization and model selection, are applicable. The methodology proposed employs a mixture of count models to cluster discrete-valued time series, in which each time series is allocated to a specific process. A simulation study is carried out, and an illustration with a real data set is made as well.

CO397 Room K0.20 ADVANCES IN ROBUST PRIOR ELICITATION Chair: Evan Kwiatkowski

C0820: **Robust external information borrowing in clinical trial hypothesis testing**
Presenter: **Silvia Calderazzo**, German Cancer Research Center (DKFZ), Germany

Bayesian clinical trials offer a natural framework for the incorporation of external information via the specification of informative prior distributions. Borrowing of such external information is often desired in order to improve the trial's efficiency and can be of crucial importance in situations where the sample size that can realistically be recruited is limited, such as paediatric or rare disease trials. An issue associated with the incorporation of external information is that external and current information may systematically differ, but such inconsistency may not be predictable or quantifiable a priori. Robust prior choices are typically proposed to avoid extreme worsening of operating characteristics in such situations. However, trade-offs in terms of frequentist characteristics are still present, and in general, no power gains are possible if strict control of type I error rate is desired. In this context, easily interpretable rationales for controlled type I error rate inflation can be of interest. An approach which allows a principled and controlled type I error rate inflation is presented. Both one and two-arm designs are considered.

C1107: **Case weighted power priors for hybrid control analyses with time-to-event data**
Presenter: **Evan Kwiatkowski**, UTHealth Houston, United States

A method is developed for hybrid analyses that use external controls to augment internal control arms in randomized controlled trials (RCT) where the degree of borrowing is determined based on the similarity between RCT and external control patients to account for systematic differences (e.g. unmeasured confounders). The method represents a novel extension of the power prior, where discounting weights are computed separately for each external control based on compatibility with the randomized control data. The discounting weights are determined using the predictive distribution for the external controls derived via the posterior distribution for time-to-event parameters estimated from the RCT. This method is applied using a proportional hazards regression model with piecewise constant baseline hazard. A simulation study and a real-data example are presented based on a completed trial in non-small cell lung cancer. It is shown that the case weighted adaptive power prior provides robust inference under various forms of incompatibility between the external controls and the RCT population.

C1113: **Eliciting prior information from clinical trials via calibrated Bayes factor**
Presenter: **Leonardo Egidi**, University of Trieste, Italy
Co-authors: Roberto Macri Demartino, Nicola Torelli, Ioannis Ntzoufras

In the Bayesian framework power, prior distributions are increasingly adopted in clinical trials and similar studies to incorporate external and past information, typically to inform the parameter associated with a treatment effect. Their use is particularly effective in scenarios with small sample sizes and where robust prior information is actually available. A crucial component of this methodology is represented by its weight parameter, which controls the volume of historical information incorporated into the current analysis. This parameter can be considered as either fixed or random. Although various strategies exist for its determination, eliciting the prior distribution of the weight parameter according to a full Bayesian approach remains a challenge. The aim is to propose a novel method for eliciting the prior distribution of the weight parameter through a simulation-based calibrated Bayes factor procedure. This approach allows for the prior distribution to be updated based on the strength of evidence provided by the data: The goal is to facilitate the integration of historical data when it aligns with current information and to limit it when discrepancies arise in terms, for instance, of prior-data conflicts. The performance of the proposed method is tested through simulation studies and applied to real data from clinical trials.

CO033 Room K0.50 COMPUTATIONAL METHODS FOR DESIGN OF EXPERIMENTS Chair: Vasiliki Koutra

C0882: **Optimal designs for state estimation in networks**
Presenter: **Kirsten Schorning**, Technical University Dortmund, Germany

The problem of estimating the expected states in networks is considered, where observations are given by repeated measurements of the random states at the nodes. The choice of the sensors directly influences the quality of the measurements at the different nodes. The problem of optimally allocating different sensors to the nodes is addressed using optimal design theory. Hereby, two models of different complexity are assumed. In the first model, the states of the different nodes are assumed to be independent of time. The design question is then used to determine which nodes need greater precision of the measurements than others. A-optimal designs are explicitly derived for different networks, e.g., a star network. In the second model, the first model is extended to time-dependent states; in particular, the states are modeled using time-dependent functions. Then, the design problem concise the optimal allocation of different measurement devices and the different time points at which measurements should be taken. Both analytical and numerical results are provided for the second model using A-optimality.

C1166: **Computationally efficient approach to operational prior specification for phase I dose-escalation trials**
Presenter: **Pavel Mozgunov**, University of Cambridge, United Kingdom

Recent years have seen increased interest in combining drug agents and schedules. Several methods for Phase I combination-escalation trials have been proposed. Most of these designs require specifying many hyper-parameters, often chosen from statistical considerations (operational prior). The conventional "grid search” calibration approach requires large simulations, which are computationally costly. A novel "cyclic calibration" has been proposed to reduce the computation from multiplicative to additive. Furthermore, calibration processes should consider a wide range of scenarios of true toxicity probabilities to avoid bias. A method to reduce scenarios based on scenario complexities is suggested. This can reduce the computation by more than 500 folds while maintaining operational characteristics similar to the grid search.

C1510: **A Laplace-based policy approach to sequential Bayesian design**
Presenter: **Emma Rowlinson**, University of Manchester, United Kingdom
Co-authors: Tim Waite

Policy-based approaches have recently developed the ability to perform sequential Bayesian designs. These approaches involve the construction of a parametric function, called the policy, mapping from the current state of knowledge to a proposed design for the next experimental run. Policy-based approaches are inherently non-myopic, formulating the policy to optimize the total expected utility over the whole sequential experiment. Optimization of the policy is achieved via stochastic gradients, which are implemented using automatic differentiation. To parametrize the policy, i.e. represent the current state of knowledge, the use of a Laplace approximation to the posterior is proposed as a compact and computationally cheap way of capturing the information amassed after each experiment. As is typical, a neural network is chosen as the policy architecture, which is trained and then can be used to inform design decisions. Initial findings when considering a linear-Gaussian example suggest the method outperforms other approaches, providing closer to optimal designs. The rationale of a Laplace parametrization is discussed, and methodology for training policies for models is developed with both continuous and discrete responses, where the latter is more challenging due to the lack of differentiability of discrete random variable simulations. The performance of the method is demonstrated through examples.

CO195 Room K2.31 (Nash Lec. Theatre) MACHINE LEARNING-BASED FAIRNESS AND CAUSAL ESTIMATION Chair: Ashkan Ertefaie

C0396: **Estimation of constrained statistical functionals for fair machine learning**
Presenter: **David Benkeser**, Emory University, United States

Constrained learning has become increasingly important, especially in the realm of algorithmic fairness and machine learning. In these settings, predictive models are developed specifically to satisfy pre-defined notions of fairness. The general problem of constrained statistical machine learning is studied through a statistical functional lens. Learning a function-valued parameter of interest is considered under the constraint that one or several pre-specified real-valued functional parameters equal zero or are otherwise bounded. The constrained functional parameter is characterized as the minimizer of a penalized risk criterion using a Lagrange multiplier formulation. Closed-form solutions for the optimal constrained parameter are often available, providing insight into mechanisms that drive fairness in predictive models. Results also suggest natural estimators of the constrained parameter that can be constructed by combining estimates of unconstrained parameters of the data-generating distribution. Thus, the estimation procedure for constructing fair machine-learning algorithms can be applied in conjunction with any statistical learning approach and off-the-shelf software. The generality of the method is demonstrated by explicitly considering a number of examples of statistical fairness constraints and implementing the approach using several popular learning approaches.

C0915: **Kernel debiased plug-in estimation**
Presenter: **Ivana Malenica**, Harvard University, United States

Modern estimation methods rely on the plug-in principle, which substitutes unknown parameters of the underlying data-generating process with estimated empirical counterparts. Flexible machine learning (ML) estimation methods have further exploited the plug-in approach. The use of highly adaptive, complex ML algorithms, however, induces plug-in bias (first-order bias) that impacts the downstream estimate. Traditional methods addressing this sub-optimal bias-variance trade-off rely on the efficient influence function (EIF) of the target parameter. When estimating multiple target parameters, these methods require debiasing the nuisance parameter multiple times using the corresponding EIFs, posing analytical and computational challenges. The targeted maximum likelihood estimation framework is leveraged to propose a novel method named kernel debiased plug-in estimation (KDPE). KDPE refines an initial estimate through regularized likelihood maximization steps, employing a nonparametric model based on reproducing kernel Hilbert spaces. It is shown that KDPE (i) simultaneously debiases all pathwise differentiable target parameters that satisfy our regularity conditions, (ii) does not require the EIF for implementation, and (iii) remains computationally tractable. The use of KDPE is numerically illustrated, and the theoretical results are validated.

C1068: **General targeted machine learning for modern causal mediation analysis**
Presenter: **Ivan Diaz**, NYU Langone Health, United States

The literature on the non-parametric definition and identification of mediational effects has grown significantly in recent years, with important progress in addressing challenges in interpreting and identifying such effects. However, statistical methodology for non-parametric estimation has lagged, with few or no methods available for tackling non-parametric estimation in continuous or high-dimensional mediators. It is shown that the identification formulas for six of the most widely known non-parametric approaches to mediation analysis proposed in recent years (natural direct and indirect effects, randomized interventional effects, separable effects, organic direct and indirect effects, recanting twin effects, and decision-theoretic effects) can be recovered from just two statistical estimands. An all-purpose, one-step estimation algorithm that can be coupled with machine learning in any mediation study that uses any of these definitions of mediation is proposed. The estimators rely on a re-parameterization of the identification formulas in terms of sequential regressions and on first-order non-parametric von Mises approximations of the first bias of a plug-in estimator to construct estimators with desirable properties, such as asymptotic normality. The one-step estimator requires the estimation of complex density ratios on the potentially high-dimensional mediators, a challenge solved using recent advancements in so-called Riesz learning.

CO102 Room K2.40 Machine learning methods and their applications in biomedical data analysis Chair: Yi Li

C0365: **Estimation and model selection for nonparametric function-on-function regression**
Presenter: **Yuedong Wang**, University of California - Santa Barbara, United States

Regression models with functional response and functional covariates have recently received significant attention. While various nonparametric and semiparametric models have been developed, there is an urgent need for model selection and diagnostic methods. A unified framework is presented for estimation and model selection in nonparametric function-on-function regression. A general nonparametric functional regression model is considered, with the model space constructed through smoothing spline analysis of variance (SS ANOVA). The proposed model reduces to some existing models when selected components in the SS ANOVA decomposition are eliminated. New estimation procedures under either L1 or L2 penalty are proposed, and combining the SS ANOVA decomposition and the L1 penalty is shown to provide powerful tools for model selection and diagnostics. Consistency and convergence rates are established for estimates of the regression function and each component in its decomposition under both the L1 and L2 penalties. Simulation studies and real examples show that the proposed methods perform well.

C1159: **Causal meta-analysis by integrating multiple observational studies with multivariate outcomes**
Presenter: **Subharup Guha**, University of Florida, United States
Co-authors: Yi Li

Integrating multiple observational studies to make unconfounded causal or descriptive comparisons of group potential outcomes in a large natural population is challenging. Moreover, retrospective cohorts, being convenience samples, are usually unrepresentative of the natural population of interest and have groups with unbalanced covariates. A general covariate-balancing framework is proposed based on pseudo-populations that extend established weighting methods to the meta-analysis of multiple retrospective cohorts with multiple groups. Additionally, by maximizing the effective sample sizes of the cohorts, a FLEXible, Optimized, and Realistic (FLEXOR) weighting method, appropriate for integrative analyses, is proposed. New weighted estimators are developed for unconfounded inferences on wide-ranging population-level features and estimands relevant to group comparisons of quantitative, categorical, or multivariate outcomes. The asymptotic properties of these estimators are examined. Through simulation studies and meta-analyses of TCGA datasets, the versatility and reliability of the proposed weighting strategy is demonstrated, especially for the FLEXOR pseudo-population.

C1117: **Causal network analysis identified mental disorders and phenotypic age acceleration as causes of dementia**
Presenter: **Hui Guo**, University of Manchester, United Kingdom

A number of biological and lifestyle factors have been associated with dementia. However, causal risk factors of the disease, which are imperative for interventions, remain elusive. Natural language processing models are utilized to select candidate risk factors of dementia from 5,505 measured variables in the UK Biobank. A holistic machine learning causal network approach is taken, fast causal inference in combination with mixed graphical models, to explore the complex causal mechanisms underlying dementia from imputed data of 120 selected variables. Of these, it is shown that eight risk factors may directly or indirectly cause dementia. The work systematically investigated causal risk factors of dementia, which paves the way for a fuller insight into its causal mechanisms. It is also shown that natural language processing models have the potential for selecting variables from high-dimensional data.

CO081 Room K2.41 Analyzing high-dimensional data with network structure Chair: George Michailidis

C0761: **Semi-parametric inference for doubly stochastic spatial point processes**
Presenter: **Ali Shojaie**, University of Washington, United States

Doubly-stochastic point processes model the occurrence of events over a spatial domain as an inhomogeneous Poisson process conditioned on the realization of a random intensity function. They are flexible tools for capturing spatial heterogeneity and dependence. However, existing implementations of doubly-stochastic spatial models are computationally demanding, often have limited theoretical guarantees, and/or rely on restrictive assumptions. A penalized regression method is presented for estimating covariate effects in doubly-stochastic point processes that are computationally efficient and do not require a parametric form or stationarity of the underlying intensity. The approach is based on an approximate (discrete and deterministic) formulation of the true (continuous and stochastic) intensity function. It is shown that consistency and asymptotic normality of the covariate effect estimates can be achieved despite the model misspecification, and develop a covariance estimator that leads to a valid, albeit conservative, statistical inference procedure. Simulation studies show the validity of our approach under less restrictive assumptions on the data-generating mechanism and an application to Seattle crime data demonstrates better prediction accuracy compared with existing alternatives.

C0863: **Joint learning of panel VAR models with low rank and sparse structure**
Presenter: **George Michailidis**, University of California, Los Angeles, United States

Panel vector auto-regressive (VAR) models are effective in modeling the evolution of multivariate time series (with an identical set of variables) across different sub-populations. The aim is to develop a panel VAR model with shared low-rank structure modulated by sub-population specific weights, enhanced by idiosyncratic sparse components. Parameter identifiability issues are addressed through constraints that lead to a nonsmooth, nonconvex optimization problem. A multiblock ADMM algorithm is developed for parameter estimation and its convergence properties established under mild regularity conditions. Further, consistency properties under high dimensional scaling are also established for the parameter estimates. The performance of the posited model is evaluated both on synthetic data and on a neuroscience data set.

C1030: **A pathwise coordinate descent algorithm for penalized quantile regression**
Presenter: **Sumanta Basu**, Cornell University, United States

A fast pathwise coordinate descent algorithm is introduced for penalized quantile regression. A closed-form update of the coordinate-wise minimization problem is derived, strategies for fast computation in high-dimension by leveraging underlying sparsity structure are discussed, and the benefit of a random perturbation is shown to help the algorithm avoid getting stuck along the regularization path. Computational efficiency gain over existing alternatives is demonstrated on simulated and real data sets.

CO236 Room S0.03 Recent advances in machine learning in econometrics Chair: Weining Wang

C1412: **Policy learning with distributional welfare**
Presenter: **Sukjin Han**, University of Bristol, United Kingdom

Optimal treatment allocation policies that target distributional welfare are explored. Most literature on treatment choice has considered utilitarian welfare based on the conditional average treatment effect (ATE). While average welfare is intuitive, it may yield undesirable allocations, especially when individuals are heterogeneous(e.g., with outliers), which is why individualized treatments were introduced in the first place. This observation motivates proposing an optimal policy that allocates the treatment based on the conditional quantile of individual treatment effects (QoTE). Depending on the choice of the quantile probability, this criterion can accommodate a policymaker who is either prudent or negligent. The challenge of identifying the QoTE lies in its requirement for knowledge of the joint distribution of the counterfactual outcomes, which is generally hard to recover even with experimental data. Therefore, minimax policies are introduced that are robust to model uncertainty. A range of identifying assumptions can be used to yield more informative policies. For stochastic and deterministic policies, the asymptotic bound is established on the regret of implementing the proposed policies. In simulations and two empirical applications, optimal decisions are compared based on the QoTE with decisions based on other criteria. The framework can be generalized to any setting where welfare is defined as a functional part of the joint distribution of potential outcomes.

C1422: **Semiparametric and nonparametric instrumental variable estimation with first-stage isotonic regression**
Presenter: **Mengshan Xu**, University of Mannheim, Germany
Co-authors: Taisuke Otsu, Kazuhiko Shinoda

A semiparametric and a nonparametric instrumental variable (IV) estimators are proposed under the assumption that the conditional mean of the endogenous variable, given the instrumental variable, is known to be monotone increasing. Isotonic estimation is employed to obtain the fitted instruments in the first stage of a two-stage semiparametric or nonparametric estimation procedure. It is shown that the proposed semiparametric IV estimator is tuning-parameter-free and achieves the semiparametric efficiency bound. Moreover, it is shown that compared to the nonparametric two-stage least squares estimator, the proposed nonparametric IV estimator requires notably fewer tuning parameters and achieves the same convergence rate. Additionally, it exhibits greater stability, as evidenced by Monte Carlo simulations.

C1549: **Change point analysis with irregular signals**
Presenter: **Tobias Kley**, Georg-August-Universitaet Goettingen, Germany
Co-authors: Yuhan Philip Liu, Hongyuan Cao, Wei Biao Wu

The problem of testing and estimating change point is considered, where signals after the change point can be highly irregular, which departs from the existing literature that assumes signals after the change point to be piece-wise constant or vary smoothly. A two-step approach is proposed to effectively estimate the location of the change point. The first step consists of a preliminary estimation of the change point that allows the obtainment of unknown parameters for the second step. In the second step, a new procedure is used to determine the position of the change point. It is shown that, under suitable conditions, the desirable $O_P(1)$ rate of convergence of the estimated change point can be obtained. The method is applied to analyze the Baidu search index of COVID-19 related symptoms, and 8 December 2019 is found to be the starting date of the COVID-19 pandemic.

CO113 Room S0.12 NEW DEVELOPMENTS IN NONPARAMETRIC STATISTICS AND NETWORK ANALYSIS Chair: Joshua Cape

C0260: **UBSea: A unified community detection framework**
Presenter: **Hao Chen**, University of California at Davis, United States
Co-authors: Xiancheng Lin

Detecting communities in networks and graphs is an important task across many disciplines, such as statistics, social science and engineering. There are generally three different kinds of mixing patterns for the case of two communities: assortative mixing, disassortative mixing and core-periphery structure. Modularity optimization is a classical way of fitting network models with communities. However, it can only deal with assortative mixing and disassortative mixing when the mixing pattern is known, and the core-periphery structure is not discovered. Modularity in a strategic way is extended, and a new framework is proposed based on unified bigroups standardized edge-count analysis (UBSea). It can address all the formerly mentioned community mixing structures. In addition, this new framework is able to automatically choose the mixing type to fit the networks. Simulation studies show that the new framework has superb performance in a wide range of settings under the stochastic block model and the degree-corrected stochastic block model. It is shown that the new approach produces a consistent estimate of the communities under a suitable signal-to-noise-ratio condition for the case of a block model with two communities for both undirected and directed networks. The new method is illustrated through applications to several real-world datasets.

C0411: **Nonparametric estimation via variance-reduced sketching**
Presenter: **Daren Wang**, University of Notre Dame, United States
Co-authors: Yuehaw Khoo, Yifan Peng

Nonparametric models are of great interest in various scientific and engineering disciplines. Classical kernel methods, while numerically robust and statistically sound in low-dimensional settings, become inadequate in higher-dimensional settings due to the curse of dimensionality. A new framework is introduced, called variance-reduced sketching (VRS), specifically designed to estimate density functions and nonparametric regression functions in higher dimensions with a reduced curse of dimensionality. The framework conceptualizes multivariable functions as infinite-size matrices and facilitates a new sketching technique motivated by numerical linear algebra literature to reduce the variance in estimation problems. The robust numerical performance of VRS is demonstrated through a series of simulated experiments and real-world data applications. Notably, VRS shows remarkable improvement over existing neural network estimators and classical kernel methods in numerous density estimation and nonparametric regression models. Additionally, theoretical justifications are offered for VRS to support its ability to deliver nonparametric estimation with a reduced curse of dimensionality.

C1240: **Euclidean mirrors and first-order changepoints in network time series**
Presenter: **Zachary Lubberts**, University of Virginia, United States

A model for a network time series is described, whose evolution is governed by an underlying stochastic process known as the latent position process, in which network evolution can be represented in Euclidean space by a curve, called the Euclidean mirror. The notion of a first-order changepoint for a time series of networks is defined, and a family of latent position process networks with underlying first-order changepoints is constructed. A spectral estimate of the associated Euclidean mirror is proven to localize these changepoints, even when the graph distribution evolves continuously but at a rate that changes. Simulated and real data examples on brain organoid networks show that this localization captures empirically significant shifts in network evolution.

CO035 Room S0.13 MODELING STRATEGIES FOR BIOMEDICAL DATA Chair: Michelle Miranda

C0575: **A fast Bayesian estimation of multi-subject fMRI activation patterns via a canonical polyadic tensor basis**
Presenter: **Michelle Miranda**, University of Victoria, Canada

Task-evoked functional magnetic resonance imaging studies, such as the Human Connectome Project (HCP), are a powerful tool for exploring how brain activity is influenced by cognitive tasks like memory retention, decision-making, and language processing. A fast Bayesian function-on-scalar model is proposed for estimating population-level activation maps linked to the working memory task. The model is based on the canonical polyadic (CP) tensor decomposition of coefficient maps obtained for each subject. This decomposition effectively yields a tensor basis capable of extracting both common features and subject-specific features from the coefficient maps. These subject-specific features, in turn, are modeled as a function of covariates of interest using a Bayesian model that accounts for the correlation of the CP-extracted features. The dimensionality reduction achieved with the tensor basis allows for a fast MCMC estimation of population-level activation maps. This model is applied to one hundred unrelated subjects from the HCP dataset, yielding significant insights into brain signatures associated with working memory.

C0595: **POI-SIMEX for conditionally Poisson distributed biomarkers from tissue microarrays**
Presenter: **Farouk Nathoo**, University of Victoria, Canada
Co-authors: Aijun Yang, Finn Hamilton, Brad Nelson, Julian Lum, Mary Lesperance

In regression analysis, covariate measurement error is an important issue that has been studied extensively. The important case is considered, where covariates are derived from tissue microarrays. In such settings, biomarkers are obtained from tissue cores that are subsampled from a larger tissue area so that these biomarkers are only estimates of the true cell densities. The resulting measurement error is non-negligible but is rarely considered in cancer studies involving tissue microarrays. These discrete biomarkers are assumed to be conditionally Poisson distributed based on a Poisson process governing the spatial locations of marker-positive cells. SIMEX is extended to the conditional Poisson case (POI-SIMEX), where measurement errors are non-Gaussian with heteroscedastic variance. The resulting POI-SIMEX estimator is shown to be strongly consistent in a linear regression model under assumptions that include a conditional Poisson distribution for the biomarker. POI-SIMEX is applied to a study of high-grade serous ovarian cancer, examining the association between survival and the presence of Tregs CD3/CD8/FOXP3 in epithelial tissue.

C0639: **Bayesian spatial model finds association between ADHD medication and long memory properties of rs-FMRI in the cerebellum**
Presenter: **Yasaman Shahhoseni**, University of Victoria, Canada
Co-authors: Farouk Nathoo, Cedric Beaulac, Michelle Miranda

Attention-deficit/hyperactivity disorder (ADHD) is a neurodevelopmental disorder common in both adults and children. Resting-state functional magnetic resonance imaging (rs-fMRI) is an important tool for investigating brain function in neurological and psychiatric disorders, including ADHD. Evidence suggests that there is an association between altered brain activity patterns and ADHD symptoms. Fractal properties of rs-fMRI time series are first explored through scale-free power spectrum properties of the brain, estimating the long-memory (LM) parameter at many locations across the brain. Further, the associations between the LM parameter and phenotypic covariates, including age and medication status of ADHD patients, are studied. A total of 140 patients with ADHD and 216 healthy controls aged 7-18 were examined. Fractal complexity is estimated using the LM parameter in the first stage, and variations are further observed in the LM maps of the brain across different individual groups. Evidence is found that participants with ADHD who were on medication exhibited a negative correlation with the LM parameter of the brain, particularly within cerebellar regions. These findings are consistent with previous neuroimaging studies of ADHD reporting associations between alterations in brain shape structure and medication in the cerebellum.

CO062 Room Safra Lec. Theatre ADVANCED TOPICS IN FUNCTIONAL AND OBJECT DATA ANALYSIS (VIRTUAL) Chair: Kuang-Yao Lee

C0502: **Geodesic mixed effects models for repeatedly observed/longitudinal random objects**
Presenter: **Satarupa Bhattacharjee**, University of Florida, United States
Co-authors: Hans-Georg Mueller

Mixed effect modelling for longitudinal data is challenging when the observed data are random objects, which are complex data-taking values in a general metric space without either global linear or local linear (Riemannian) structure. In such settings, the classical additive error model and distributional assumptions are unattainable. Due to the rapid advancement of technology, longitudinal data containing complex random objects, such as covariance matrices, data on Riemannian manifolds, and probability distributions, are becoming more common. Addressing this challenge, a mixed-effects regression is developed for data in geodesic spaces, where the underlying mean response trajectories are geodesics in the metric space, and the deviations of the observations from the model are quantified by perturbation maps or transports. A key finding is that the geodesic trajectories assumption for the case of random objects is a natural extension of the linearity assumption in the standard Euclidean scenario to the case of general geodesic metric spaces. Geodesics can be recovered from noisy observations by exploiting a connection between the geodesic path and the path obtained by global Frachet regression for random objects. The effect of baseline Euclidean covariates on the geodesic paths are modelled by another Frachet regression step. The asymptotic convergence of the proposed estimates is studied, and illustrations are provided through simulations and real-data applications.

C0589: **Deep Frechet regression**
Presenter: **Yidong Zhou**, University of California, Davis, United States
Co-authors: Su I Iao, Hans-Georg Mueller

Advancements in modern science have led to the increasing availability of non-Euclidean data in metric spaces. The challenge of modeling relationships is addressed between non-Euclidean responses and multivariate Euclidean predictors. A flexible regression model is proposed, capable of handling high-dimensional predictors without imposing parametric assumptions. Two primary challenges are addressed: the curse of dimensionality in nonparametric regression and the absence of linear structure in general metric spaces. The former is tackled using deep neural networks, while for the latter, the feasibility of mapping the metric space is demonstrated where responses reside to a low-dimensional Euclidean space using manifold learning. A reverse mapping approach is introduced, employing local Frechet regression to map the low-dimensional manifold representations back to objects in the original metric space. A theoretical framework is developed, investigating the convergence rate of deep neural networks under dependent sub-Gaussian noise with bias. The convergence rate of the proposed regression model is then obtained by expanding the scope of local Frechet regression to accommodate multivariate predictors in the presence of errors in predictors. Simulations and case studies show that the proposed model outperforms existing methods for non-Euclidean responses, focusing on the special cases of probability measures and networks.

C1261: **Structure-preserving nonlinear sufficient dimension reduction for scalar-on-tensor regression**
Presenter: **Dianjun Lin**, Pennsylvania State University, United States
Co-authors: Lingzhou Xue, Bing Li

A novel approach is presented to nonlinear sufficient dimension reduction for scalar-on-tensor regression and classification problems. The method introduces the tensor envelope, a framework designed to preserve the intrinsic multidimensional structure of tensor-valued predictors while achieving effective dimension reduction. Additionally, the central dimension folding subspace is defined within which the tensor envelope acts as an operator. Using coordinate mapping, two optimization algorithms are developed to enhance the operators' objective function. The performance of the proposed estimators is assessed through comprehensive simulations and real-world applications.

CO198 Room BH (S) 1.01 Lec. Theathre 1 NONLINEAR AND NON-GAUSSIAN TIME SERIES Chair: Sean Telg

C1290: **A panel extension for noncausal models**
Presenter: **Kevin Cecere Palazzo**, Vrije Universiteit, Netherlands
Co-authors: Sean Telg, Siem Jan Koopman, Francisco Blasques

Noncausal autoregressive models offer a useful framework to model speculative bubbles in economics and finance, thanks to their ability to generate explosive phenomena. Standard literature on common features for multivariate noncausal models implicitly assumes that common features occur at the same time. A panel extension of noncausal models is considered in order to accommodate for joint modeling of financial bubbles when explosive behaviors are not synchronous across units. Parameters are proposed for estimation by approximate maximum likelihood (AML) and least squares (LS), and the asymptotic behavior of the proposed estimators is established. Properties of limit behavior of the proposed estimators are analyzed in finite and infinite variance frameworks. A testing procedure is derived to test the validity of panel extension in noncausal framework, and the potential usefulness of the results is illustrated in an empirical application on cryptocurrencies.

C1571: **A novel test for the presence of local explosive dynamics**
Presenter: **Sean Telg**, Vrije Universiteit Amsterdam, Netherlands
Co-authors: Siem Jan Koopman, Francisco Blasques, Gabriele Mingoli

In economics and finance, speculative bubbles take the form of locally explosive dynamics that eventually collapse. A test for the presence of speculative bubbles is proposed in the context of mixed causal-noncausal autoregressive processes. The test exploits the fact that bubbles are anticipative; they are generated by an extreme shock in the forward-looking dynamics. In particular, the test uses both path level deviations and growth rates to assess the presence of a bubble of a given duration and size at any moment of time. It is shown that the distribution of the test statistic can be either analytically determined or numerically approximated, depending on the error distribution. The size and power properties of the test are analyzed in controlled Monte Carlo experiments. An empirical application is presented for a monthly oil price index. It demonstrates the ability of the test to detect bubbles and to provide valuable insights in terms of risk assessments in the spirit of Value-at-Risk.

C1451: **Nonparametric time-varying Granger causality using exponentially smoothed density estimators**
Presenter: **Sicco Kooiker**, Vrije Universiteit, Netherlands

Where parametric methods require assumptions about the unknown type of Granger causality, and static methods tend to over-reject and lack power in dynamically changing time series environments, the nonparametric time-varying Granger causality (NPTVGC) testing framework proves to be a useful method. A procedure is presented that combines the popular Diks-Panchenko (DP) test with exponentially weighted moving average local density estimators to assess Granger causality in environments that change gradually over time. Originally, DP suggested setting the bandwidth using their rule-of-thumb. This method becomes invalid under smoothing. A cross-validation hyperparameter optimization algorithm is introduced, providing an alternative method for selecting the bandwidth of the DP statistic. Two simulation studies demonstrate the importance of correct hyperparameters and the use of exponential weighting. In the empirical experiment, the NPTVGC framework identifies Granger causality from the Hang Seng stock index to the KOSPI stock index at times when traditional linear Granger causality methods do not. This demonstrates that the framework is beneficial when the functional form of Granger causality is misspecified.

CO038 Room BH (SE) 1.01 BAYESIAN METHIDS FOR RECORD LINKAGE AND SMALL AREA ESTIMATION Chair: Jairo Fuquene

C1285: **Genealogical application of record linkage for black Americans in the Antebellum South**
Presenter: **Hannah Butler**, Colorado State University, United States
Co-authors: Andee Kaplan

Record linkage is used to connect records that come from the same entity across multiple data sources. Probabilistic record linkage is utilized without reliable identifying information to estimate probabilities that two records refer to the same entity. Entities recognized by alternate information in different contexts can manifest as multiple distinct records for a single entity appearing within or between different data sources. An alias is the occurrence of one or more duplications of an entity within a file, not due to error but rather due to a known alternative piece of information. Aliases are separate parts of the story and can provide richer data to link records. However, data containing aliases requires a more careful approach to statistical inference. In existing record linkage methodologies, pre- or post-hoc processing may be done to avoid or remove conflicting links due to aliases. This has the consequence of losing potentially valuable information or impairing the ability to quantify uncertainty. A fully Bayesian approach is proposed to record linkage that expands the existing methodology to account for and leverage known aliases of entities within data files to be linked. This approach also allows for uncertainty quantification and requires no post-hoc processing of link estimation. The performance of this approach is demonstrated in simulation, and the model is applied to two sources of data from Freedom-Seekers in the Antebellum South.

C1280: **A hierarchical gamma prior for modeling random effects in small area estimation**
Presenter: **Xueying Tang**, University of Arizona, United States

Small area estimation (SAE) is becoming increasingly popular among survey statisticians. Since the direct estimates of small areas usually have large standard errors, model-based approaches are often adopted to borrow strength across areas. SAE models often use covariates to link different areas and random effects to account for the additional variation. Recent studies showed that random effects are not necessary for all areas, so global-local (GL) shrinkage priors have been introduced to effectively model the sparsity in random effects. The GL priors vary in tail behavior, and their performance differs under different sparsity levels of random effects. As a result, one needs to fit the model with different choices of priors and then select the most appropriate one based on the deviance information criterion or other evaluation metrics. A flexible prior is proposed for modeling random effects in SAE. The hyperparameters of the prior determine the tail behavior and can be estimated in a fully Bayesian framework. Therefore, the resulting model is adaptive to the sparsity level of random effects without repetitive fitting. The performance of the proposed prior is demonstrated via simulations and real applications.

C1516: **Bayesian alternatives to model the variances of direct estimates**
Presenter: **Sirapat Watakajaturaphon**, University of California, Davis, United States

The estimates of variances in small area estimation play an important role. Common problems of assuming a frequentist framework for modeling the variances of small area estimates are discussed and a new Bayesian framework is proposed to deal with these problems in practice. Suitable Markov chain Monte Carlo algorithms are proposed, and the theoretical properties of the proposed model is studied. Finally, the model in a real data set is implemented.

CO109 Room BH (SE) 1.02 BAYESIAN SEMI- AND NON-PARAMETRIC METHODS I Chair: Andrea Cremaschi

C0409: **A Bayesian nonparametric approach for nonignorable missingness in EHR data**
Presenter: **Michael Daniels**, University of Florida, United States
Co-authors: Sebastien Haneuse, David Lindberg

An approach for missingness in EHRs is proposed using Bayesian nonparametric (BNP) models. It shows how to introduce sensitivity parameters corresponding to nonignorable missingness in the outcome and confounders by extracting unidentified distributions from the BNP model and reconstructing the distribution of interest. Auxiliary covariates are also flexibly included to move closer to MAR. G-computation is used based on the reconstructed distribution to compute causal estimands of interest.

C0637: **Bayesian inference via predictive distributions**
Presenter: **Samuele Garelli**, University of Bologna, Italy

Predictive distributions offer an appealing alternative to the likelihood-prior-posterior approach to Bayesian inference. In fact, if predictive distributions have a good fit on the observed data and converge in some sense, they can be used to perform inference. Moreover, choosing $P(X_{n+1}|X_1,...,X_n)$ is more natural than specifying a prior distribution since the former is defined on the data (i.e. observables) while the latter is defined on parameters (i.e. unobservables). In practice, inference can be carried out by reconstructing the unobserved part of the population via recursive sampling from the predictive distributions and by taking statistics of the observed and the imputed data together. A way to define predictive distributions that are both theoretically and computationally tractable is via mixtures of distributions initialized by a clustering algorithm whose dynamics are driven by the mean and the variance of each cluster. Such predictive distributions enjoy interesting convergence properties and can be used to target three main inferential tasks, i.e. parameter estimation, regression and classification.

C1691: **A phylogenetic model of the evolution of discrete matrices for the inference of lexical & phonological language history**
Presenter: **Robin Ryder**, Imperial College London, United Kingdom

A model of the evolution of a matrix along a phylogenetic tree is introduced, in which transformations affect either entire rows or columns of the matrix. This represents the change of both lexical and phonological aspects of linguistic data, by allowing for new words to appear and for systematic phonological changes to affect the entire vocabulary. We implement a Sequential Monte Carlo method to sample from the posterior distribution, and infer jointly the phylogeny, model parameters, and latent variables representing cognate births and phonological transformations.

We successfully apply this method to synthetic and real data of moderate size.

CO211 Room BH (SE) 1.05 MACHINE LEARNING AND BAYESIAN METHODS IN FINANCE Chair: Martina Zaharieva

C0659: **Modelling extreme joint dependence using Bayesian nonparametric copulas**
Presenter: **Concepcion Ausin**, Universidad Carlos III de Madrid, Spain
Co-authors: Maria Kalli

Modelling the joint distribution of financial variables is not an easy task. A flexible approach is needed in order to capture both the tail and central dependence structure as well as the slight asymmetry of the distribution. Copulas are useful in capturing the joint dependence of multiple random variables. However, parametric copulas are not flexible enough to capture both tail and central dependence, especially in periods of financial crises when tail dependence is higher and usually more important than overall correlations. From the Bayesian point of view, only a few nonparametric copula models have been proposed, and none of them adequately account for tail dependencies. A new Bayesian nonparametric copula is proposed, which can be viewed as a multivariate histogram smoothing with a non-equally spaced infinite number of bins. A prior is imposed on the breakpoints and on the volume of the bins. It is shown how to express the model as an infinite mixture of beta distributions using the stick-breaking representation of the Dirichlet process. Inference and prediction are made using a Gibbs sampler. The procedure is illustrated using simulated and real data based on multivariate financial time series.

C0898: **Mutually dependent Bernoulli processes for multivariate change-point detection**
Presenter: **Carson McKee**, Kings College London, United Kingdom
Co-authors: Maria Kalli

Financial and economic time series exhibit sharp structural breaks, or change-points, driven by events such as recessions or financial market crashes. In the multivariate setting, these changes may occur asynchronously across series. It is often the case that the occurrence of a change-point in one series affects the probability of a subsequent change occurring in another. This is observed, for example, in financial contagion, when turmoil spreads from one country to another. A multivariate change-point prior is developed, which explicitly models this dependence using discrete time and mutually dependent point processes. Under this prior, the probability of a change-point occurring in a series at a given time is dependent on recent changes in other series. Thus, the model allows for both cross-sectional and temporal dependence in the change-point probabilities. Then, conditional on the change-point locations, the data in each segment is assumed to be independent of the data in other segments. Obtaining posterior estimates under this model is non-trivial. A blocked Gibbs sampler and a particle Gibbs sampler are developed for use in high dimensions. The model is demonstrated on simulated and real datasets and is shown to uncover latent dependencies linking change-points in different series.

C0964: **What events matter for exchange rate volatility**
Presenter: **Igor Martins**, Orebro University, Sweden

The purpose is to expand stochastic volatility models by proposing a data-driven method to select the macroeconomic events most likely to impact volatility. The effect of macroeconomic events on exchange rate volatility in multiple countries is identified and quantified in multiple countries using high-frequency currency returns while accounting for persistent stochastic volatility effects and seasonal components that capture time-of-day patterns. Due to the hundreds of macroeconomic announcements and their lags, sparsity-based methods are relied on to select relevant events for the model. The contribution to the exchange rate literature is in four ways. First, the macroeconomic events that drive currency volatility are identified, their effect is estimated and connected to macroeconomic fundamentals, and how they can be linked to lower-frequency currency returns is shown using a model averaging argument. Second, a connection between intraday seasonality, trading volume, and opening hours of major markets across the globe is found, and a simple labor-based argument is provided for the pattern found. Third, it is shown that the inclusion of macroeconomic events and seasonal components is key for forecasting exchange rate volatility. Fourth, applying the proposed model for multiple currencies alongside a dynamic copula yields a Sharpe ratio 3.5 times higher than using standard SV and GARCH models.

CO366 Room BH (SE) 1.06 RECENT ADVANCES ON NONPARAMETRIC PANEL DATA ANALYSIS Chair: Juan Manuel Rodriguez-Poo

C0383: **Labor income tax shocks and corporate innovation**
Presenter: **Daniel Henderson**, University of Alabama, United States
Co-authors: Alexandra Soberon, Taining Wang, Soroush Ghazi

Narrative identification and local projection are used to study the effect of income tax shocks on innovation measures of publicly traded firms in the US. To relax conditions on the way in which exogenous tax shocks impact our innovation measures, a semiparametric panel model is developed to estimate trending coefficient functions in the presence of interactive fixed effects. A kernel-based profile estimator is employed, and a local linear approach is used to filter the unobserved cross-sectional factors in the presence of mixed covariates (i.e., continuous and discrete). The relevant asymptotic theory is developed, and the estimators are shown to perform well in finite samples. In the application, on average, it is shown that a one percentage point cut in the average marginal tax rate increases firms' R&D expenditure by about one per cent after four years, but that the impact is heterogenous with respect to whether or not the firm has a patent, the intensity of R&D expenditure and the sign of the tax shock.

C0405: **Estimation of functional coefficient panel data models with sample selection and fixed effects: A pairwise approach**
Presenter: **Alexandra Soberon**, Universidad de Cantabria, Spain
Co-authors: Juan Manuel Rodriguez-Poo, Daniel Henderson, Taining Wang

The focus is on the consistent estimation of a flexible functional-coefficient panel data sample selection model with fixed effects that enables the capture of the potential parameter heterogeneity in the relationship of interest. A two-step estimation procedure is proposed that avoids identification restrictions based on a pairwise transformation. The first stage estimates the unknown parameters of the selection equation consistently, while the second stage estimates the regression of interest using these estimates and a generalized local weighting matrix that enables the removal asymptotically of the sample selection bias. The asymptotic distribution of the proposed estimators is analyzed under rather weak assumptions. Simulations reveal an excellent performance of the proposed estimators.

C0487: **Estimation and inference of panel data models with a generalized factor structure**
Presenter: **Stefan Sperlich**, University of Geneva, Switzerland
Co-authors: Juan Manuel Rodriguez-Poo, Alexandra Soberon

A novel panel data model is introduced with a fairly general structure for the unobserved common factors, which also encompasses both traditional additive and interactive fixed effects. Under rather weak assumptions, consistent estimators are obtained that are asymptotically normal at rate root-NT for the parameters of interest, while optimal nonparametric estimators are obtained for the unspecified part. Furthermore, the statistical properties of the resulting estimators are robust to misspecification of the relationship between common factors and factor loadings. A nonparametric specification test for the crucial modeling assumption is provided. It relies on combining the methodology of conditional moment tests and nonparametric estimation techniques. Using degenerate and nondegenerate theories of U-statistics, its convergence and asymptotic distribution are shown under the null, and it diverges under the alternative at a rate arbitrarily close to root-NT. Finite sample inference is based on bootstrap. Simulations reveal an excellent performance of the methods. They are used to study the effect of the European Union emissions trading system on CO2 emission and the economy of some countries of the European Union.

C0515: **An empirical likelihood goodness-of-fit test for panel data models with interactive fixed effects**
Presenter: **Luis Antonio Arteaga Molina**, Universidad de Cantabria, Spain
Co-authors: Juan Manuel Rodriguez-Poo

The empirical likelihood device for a panel data model is employed with interactive fixed effects to formulate a test statistic that measures the goodness of fit of a parametric regression model. The asymptotic distribution of the test statistic is derived, and a Bootstrap procedure is also proposed to obtain the critical values. The technique is based on a comparison with kernel smoothing estimators. The empirical likelihood formulation of the test has two attractive features. One is its automatic consideration of the variation that is associated with the nonparametric fit due to empirical likelihood's ability to Studentize internally. The other is that the asymptotic distribution of the test statistic is free of unknown parameters, avoiding plug-in estimation. To show the feasibility of the technique and to analyze its small sample properties, a Monte Carlo simulation exercise is implemented, and the proposed technique is also illustrated in an empirical analysis of the environmental Kuznets curve hypothesis.

CO145 Room BH (S) 2.03 TOPICS IN FINANCIAL ECONOMETRICS Chair: Florian Richard

C1011: **The economic value of reward-to-risk timing strategies using return-decomposition GARCH models**
Presenter: **Arsene Brou**, Laval University, Canada
Co-authors: Richard Luger

In portfolio management, reward-to-risk timing strategies require estimates of expected returns in addition to volatility estimates. To address this need, a new GARCH-type model is proposed based on a decomposition of returns into their signs and absolute values. The conditional volatility is determined by innovations following a folded normal distribution, and the conditional mean depends on the skewness dynamics implied by the interaction between the multiplicative sign and absolute return components. The out-of-sample performance of this approach is compared with the naive diversification rule, the plug-in approach, and other GARCH-type specifications. The empirical analysis of daily stock returns demonstrates the economic value of exploiting the implied time-varying skewness for reward-to-risk timing strategies.

C1423: **Monetary policy surprises: Robust dynamic causal effects**
Presenter: **Haowei Tang**, Carleton University, Canada
Co-authors: Lynda Khalaf

The one-step local projection instrumental variable (LP-IV) approach is built on assessing the effects of a monetary policy shock on the US economy, focusing on inference. Methodologically, particular attention is paid to the specification of control variables that warrant exogeneity of the instrumental variables. A response parameter is introduced that is always identified and can account for misspecification. This response parameter enables us to rely on least-squares even when endogeneity is present. Associated findings are robust in expanding credit spread data and in various competing instruments that reflect conflicting views about the nature of central bank communication and the informational advantage of central banks.

C1193: **Simulation-based multiple testing for many non-nested multivariate models**
Presenter: **Florian Richard**, Universite Laval, Canada
Co-authors: Lynda Khalaf

A multivariate extension of exact specification tests for non-nested models is proposed. The test is finite-sample exact under the assumption of Gaussian errors and is easily generalized to a multiple-model hypothesis via a combined alternative. Valid inference results are obtained using bootstrapped Monte Carlo p-values, even when the distribution under the null hypothesis is intractable. Both Gaussian and non-Gaussian error structures are considered through bootstrapping, and it is shown that the test possesses good size and power properties via simulations. Finally, empirical applications to asset pricing are presented by testing benchmark factor models against single and multiple alternatives.

CO210 Room BH (S) 2.05 STOCHASTIC DOMINANCE AND APPLICATIONS IN FINANCE Chair: Nikolas Topaloglou

C1105: **Are commodity markets segmented: Understanding cross-asset interdependencies using stochastic spanning**
Presenter: **Argyro Kofina**, Athens University of Economics and Business, Greece
Co-authors: Evgenia Passari, Nikolas Topaloglou

There is mixed evidence on the integration of commodity markets with equity and bond markets. On the one hand, a number of papers document that standard equity asset pricing factors cannot explain the cross-section of commodity futures returns, implying market segmentation. On the other hand, there is some evidence that the recent financialization of commodities tends to integrate markets. A unified approach is proposed for the study of commodity price behaviour that builds on a existing theory. According to the theory, large supply shocks exacerbate market segmentation by limiting the willingness of market specialists to trade across markets. At the same time, it is explored whether demand shocks encourage market integration through increased investing across asset classes and during periods of big financial distress as investors redirect portfolio flows out of risky portfolios and into safe assets. In particular, the differential degree of market segmentation is formally tested for under narratively identified supply and demand commodity developments by employing a non-parametric, distribution-free measure of stochastic dominance.

C1181: **On the existence of a true mutual fund factor model**
Presenter: **Nikolas Topaloglou**, Athens University of Economics and Business Research Center, Greece
Co-authors: Argyro Kofina, Ioannis Psaradelis

A mutual fund factor model is introduced from a large set of 168 factors, including characteristics found to be significant in explaining stock returns, leading factors explaining mutual fund returns and macroeconomic variables. To do so, a stochastic dominance approach is used that extends the mean-variance framework underlying the usual APT tests of exact pricing to include fairly general sets of preferences for the arbitrageurs operating in the economy. The proposed factor model explains mutual fund returns better than all leading factors and a Lasso-based factor model when using both parametric, non-parametric, and machine learning tests. The results are significant both in-sample and out-of-sample.

C1223: **ETFs, stochastic dominance and market efficiency**
Presenter: **Ioannis Psaradellis**, University of Edinburgh, United Kingdom
Co-authors: Nikolas Topaloglou

The market efficiency of active mutual funds and ETFs is investigated relative to their stock ownership using a stochastic dominance approach that extends the mean-variance framework. In particular, stochastic bound portfolios are constructed from fund holdings, and those portfolios' dominance is assessed against the reference portfolio set of the traded mutual funds and ETFs. The improvement in investment performance and enhanced benchmarking are examined using funds' reference sets and stochastic bounds on fund ownership. Empirical results using both parametric and non-parametric tests indicate that the stochastic bound portfolios dominate the reference portfolio set. The results are significant both in-sample and out-of-sample.

CO006 Room BH (SE) 2.01 PROJECTION PURSUIT Chair: Nicola Loperfido

C0560: **Bayesian projection pursuit for efficient and sustainable banking**
Presenter: **Alessandro Berti**, Urbino University Carlo Bo, Italy

Co-authors: Nicola Loperfido, Cinzia Franceschini

The European banking agency (EBA) established new and cogent guidelines (EBA-LOM) that European banks should follow when granting credit. The guidelines were set in June 2021 but have been implemented in the current year. There is a general agreement that their implementation will be gradual. The main novelty of EBA-LOM guidelines, in addition to providing precise and well-defined indications on tools for credit risk measurement, is their emphasis on environment, sustainability and governance factors (ESG factors), which are non-financial factors. For example, banks should be careful when granting credit to firms with questionable attitudes towards their workers. The EBA-LOM guidelines pose several problems to banks, which did not occur when credit is granted using balance sheet ratios only. Firstly, how should ESG factors be measured? Secondly, how can ESG data be collected from firms? Thirdly, how should these data be summarized into a single measure of credit risk? The default approach relies on software prepared by rating agencies (Moody's, Fitch, and Standard & Poor's), which do not consider the unique characteristics of given firms. A Bayesian approach is proposed based on prior elicitation of ESG factors given by professional bank consultants, which are then merged with other information by means of projection pursuit. The approach is illustrated with a small dataset of Italian firms.

C0609: **Projection pursuit for art analysis**
Presenter: **Cinzia Franceschini**, University of Sassari, Italy
Co-authors: Nicola Loperfido, Noemi Loperfido

"Composition II in Red, Blue and Yellow" by Piet Mondrian is an abstract painting composed of blue, red, white and yellow rectangles separated by thick, black, orthogonal lines. It is the most iconic painting by Piet Mondrian, and it inspired product design, communication design, interior design, and fashion design. The painting has been extensively studied using both formal and historical approaches. Most art experts agree that the painting has a dynamic equilibrium with deep aesthetic merits. These beliefs are investigated by collecting the opinions of teenagers attending the last year of an art school. Data were analyzed using several multivariate techniques, including principal component analysis and projection pursuit. The latter outperformed the former in finding data structures, for example, clusters and outliers. The results encourage the use of statistical methods when analyzing works of art.

C0872: **Projection pursuit and portfolio selection**
Presenter: **Chris Adcock**, Sheffield University Management School, United Kingdom

Projection pursuit is a data analysis technique that generates one-dimensional sum-maries of complex multivariate data sets. The methods operate on estimates of themultivariate moments of the original n-dimensional data. For variance projection pursuit based on a covariance matrix Σ, the method computes the values of an n-vector w that maximizes the quadratic form $w^T \Sigma w$ subject to the normalization $w^T w = 1$. In portfolio theory, the data set comprises returns on a set of risky financial assets, and the elements of the vector w denote investment proportions. The minimum variance portfolio is computed by minimizing the same quadratic form. In this case, the normalization employed is that the weights sum to unity. It is usually also the case that the weight is required to be non-negative. The use of the same quadratic form suggests that the normalization used in portfolio theory could also be employed as an alternative method of data reduction in projection pursuit. In addition, the pursuit of skewness projection suggests variations in portfolio theory methods. Standard mean-variance portfolio selection could be replaced by skewness-variance or other combinations of higher moments. The aim is to investigate differences as well as similarities between projection pursuit and portfolio selection. Both potential synergies and aspects are reported, where the two methods are distinct.

CO118 Room BH (SE) 2.05 NEW CHALLENGES FOR STATISTICAL PROCESS CONTROL Chair: Claudio Giovanni Borroni

C1060: **Control charts for dynamic process monitoring with an application to air pollution surveillance**
Presenter: **Peihua Qiu**, University of Florida, United States

Air pollution is a major global public health risk factor. To tackle problems caused by air pollution, governments have put a huge amount of resources into improving air quality and reducing the impact of air pollution on public health. In this effort, it is extremely important to develop an air pollution surveillance system to constantly monitor the air quality over time and give a prompt signal once the air quality is found to deteriorate so that a timely government intervention can be implemented. To monitor a sequential process, a major statistical tool is the statistical process control (SPC) chart. However, traditional SPC charts are based on the assumption that process observations at different time points are independent and identically distributed. These assumptions are rarely valid in environmental data because seasonality and serial correlation are common in such data. To overcome this difficulty, a new control chart is suggested that can properly accommodate dynamic temporal patterns and serial correlation in a sequential process. Thus, it can be used for effective air pollution surveillance.

C0777: **Comparing the treatment of old data in some self-starting control charts**
Presenter: **Manuela Cazzaro**, University of Milano-Bicocca, Italy
Co-authors: Claudio Giovanni Borroni, Paola Maddalena Chiodini

Self-starting control charts do not distinguish between Phase I and Phase II in the statistical control task. Nonetheless, control cannot really start unless a minimum number of readings is collected and, according to what is reported by many authors, such minimum needs often to be raised to get a suitable performance. That depends mostly on how the chart tentatively separates between the new and the old observations and on how it is instructed to learn from the second kind of data. Being that, in some applications, the availably of a large reference sample (i.e. collected under normal conditions) is often prevented, the aim is to compare the strengths of some recent proposals when the control needs to be started after very few initial readings are gathered. The analysis is based on some simulations from distributions with different shapes; in addition, it considers different kinds of shifts from the in-control to the out-of-control situation and different instants when that shift occurs after control has started.

C1263: **Extending a novel class of control charts for sequential monitoring to the multivariate framework**
Presenter: **Claudio Giovanni Borroni**, University of Milano - Bicocca, Italy
Co-authors: Manuela Cazzaro

The change-point paradigm has been successfully applied to statistical process control by building many parametric and nonparametric charts aimed at the sequential monitoring of some univariate process characteristics. It is known, however, that the quality of a process is often measured by multivariate variables and also that the dependence structure of such variables cannot be ignored to provide a prompt signal of the drift of the process from the in-control situation. In the extension of univariate control tools to the multivariate case, the so-called interpoint distances play a key role and can be variously defined. A suitable version of such distances is sought to fit a non-conventional approach to the change-point methodology, which has been studied in the univariate setting so far. The focus is on the efficacy of the resulting multivariate control charts but also on the simplicity of their application.

CO009 Room BH (SE) 2.09 APPLIED MACRO-FINANCE II Chair: Alessia Paccagnini

C1531: **The economic superpower: Analyzing the impact of superhero movies on the U.S. business cycle**
Presenter: **Alessia Paccagnini**, University College Dublin, Ireland

The relationship between the growing popularity of superhero movies and the U.S. business cycle is investigated. Building on the observation that superheroes gain popularity during economic downturns, quarterly U.S. box office revenues are analyzed from 1978 to 2023 alongside key economic indicators such as GDP, inflation, unemployment, and monetary policy. Using a proxy-structural vector autoregression (Proxy-SVAR) model and a novel movie-based variable as a proxy for uncertainty, the findings suggest that the rise in superhero movies may have negative impacts on macroeconomic indicators, comparable to traditional uncertainty shocks like the VIX and the economic policy uncertainty index.

C1599: **How does the US stock market react to climate concern shocks across frequencies: A wavelet analysis**
Presenter: **Andrea Cipollini**, University of Palermo, Italy
Co-authors: Iolanda Lo Cascio, Fabio Parla, Fabio Parla

A number of studies on climate risk hedging through investment in equities have examined the performance of a green-minus-brown (GMB) portfolio of stocks in terms of both the expected and unexpected portfolio stock return performance. The recent study explains the observed outperformance of green stocks relative to brown stocks (for the U.S.) through the unexpected component of the GMB portfolio return, driven by innovation to a climate concern index. It is argued that another study assumes that the role played by climate concern is invariant to frequency. In line with another study, a spectral factor model is used to study whether systematic risk, in particular climate risk, varies across frequencies, hence whether a rise in climate concern develops especially over medium-term cycles (between 16 and 32 months) rather than short-term ones (between two and four months). The analysis is carried out by controlling for traditional Fama French factors and employing a factor decomposition of the covariance matrix of wavelet coefficients of the financial time series used using the MODWT filter.

C1643: **The distributional effects of stabilization policy**
Presenter: **Laura Jackson Young**, Bentley University, United States
Co-authors: Michael Owyang, Alessia Paccagnini

Recent studies focus on the effects of monetary policy on the distribution of income. Generally, expansionary monetary policy (i.e., lower interest rates) is thought to increase income inequality via capital gains income. It is argued that any expansionary stabilization policy -monetary or fiscal-has the same distributional effects. Even policies that target income inequality, such as distributional tax policy, can raise income inequality if their net effects are expansionary.

CO124 Room BH (SE) 2.10 RECENT ADVANCEMENTS IN MODERN MULTIVARIATE PROBLEMS Chair: Chenlu Ke

C0825: **Dimension reduction for spatially correlated data**
Presenter: **Hossein Moradi Rekabdar.**, South Dakota State University, United States

Dimension reduction provides a useful tool for statistical data analysis with high-dimensional data. A parsimonious multivariate spatial regression model is developed with a non-separable covariance function. The efficacy of this new solution is illustrated through simulation studies and real data analysis. It is shown that for cases where the marginal spatial correlations are different from each other, the proposed non-separable model provides better estimation and inference than the related separable model and provides tighter inference than a non-separable spatial model without dimension reduction when there is immaterial variation in the data.

C1220: **Multivariate rank-based expectation of the conditional difference for testing independence**
Presenter: **Xiaoli Kong**, Wayne State University, United States

The purpose is to introduce a class of multivariate rank-based measures for testing independence, which utilize the expected conditional characteristic function-based independence criterion. The tests are computationally efficient and remain well-defined under minimal assumptions on the underlying distributions. The asymptotic distribution for these test statistics is established, and their effectiveness is demonstrated through simulation studies. Additionally, the practical utility of the proposed test is illustrated through a real data application.

C1079: **Some modeling considerations involving the exponentially-modified Gaussian (EMG) distribution**
Presenter: **Yanxi Li**, Metropolitan State University of Denver, United States

Fitts' law is often employed as a predictive model for human movement, especially in the field of human-computer interaction. Models with an assumed Gaussian error structure are usually adequate when applied to data collected from controlled studies. However, observational data (often referred to as data gathered "in the wild") typically display noticeable positive skewness relative to a mean trend as users do not routinely try to minimize their task completion time. As such, the exponentially modified Gaussian (EMG) regression model has been applied to aimed movement data. However, it is also of interest to reasonably characterize those regions where a user likely was not trying to minimize their task completion time. A novel model is proposed with a two-component mixture structure - one Gaussian and one exponential - on the errors to identify such a region. An expectation-conditional-maximization (ECM) algorithm is developed to estimate such a model, and some of the algorithm's properties are established. The efficacy of the proposed model, as well as its ability to inform model-based clustering, are addressed through extensive simulations and an insightful analysis of a human-aiming performance study.

CO090 Room BH (SE) 2.12 STATISTICAL METHODS FOR ANALYZING BIG AND COMPLEX DATA (VIRTUAL) Chair: Trambak Banerjee

C0355: **Statistical inference for subgraph densities under induced random sampling from network data**
Presenter: **Nilanjan Chakraborty**, Missouri University of Science and Technology, United States
Co-authors: Ayoushman Bhattacharya, Soumen Lahiri

Statistical inference for large networks based on sampled smaller network data is an important problem in network analysis. The focus is on developing a framework for obtaining statistical guarantees for subgraph densities of a general population network under without replacement sampling (SRSWOR). Examples of such subgraph densities include edge density, triangle density, two-star density and other popularly studied graph summary statistics. Under this sampling scheme, a Berry-Esseen bound is derived to establish the asymptotic normality of the Horwitz-Thompson (HT) estimator for the population subgraph densities. The HT estimator is shown to be unbiased for population subgraph densities. To facilitate inferential procedures, a jackknife estimator of the unknown population variance is provided, and its consistency is established. The joint asymptotic normality of two subgraph densities is also established, which is crucial in establishing the asymptotic normality of the global clustering coefficient/global transitivity of the sampled graph. Results find a useful application to the problem of testing the equality of two population graphs using the subgraph densities as the test statistic. Finally, a simulation study is presented, which corroborates the theoretical findings.

C1467: **Detecting structural changes in time varying parameters of panel models**
Presenter: **Padma Sharma**, Federal Reserve Bank of Kansas City, United States

A hierarchical Bayesian procedure is developed to study the dynamics of bank stock returns to changes in their capacity to provide liquidity as well as the strength of their capital position and identify structural changes in this relationship over the last 30 years. The hierarchical model relies on a dynamic extension of the spike-and-slab prior that identifies change points in the relationship between bank stock prices and their liquidity buffers as well as their capital ratios. The proposed framework detects distinct structural changes across different covariates, which remain undetected by existing methods that only detect a set of common change points across all covariates. The analysis uncovers previously overlooked instances of structural changes in the relationship between bank liquidity and equity returns. Structural changes occur less frequently in the relationship between capital ratio and equity returns and are limited to periods of severe market distress, such as the global financial crisis and the onset of COVID-19. The method highlights the importance of a flexible method for detecting structural changes that allow for different instances of change points across covariates.

C1470: **Latent group structures and sparsity analysis in high dimensional panel MIDAS models**
Presenter: **Shahnaz Parsaeian**, University of Kansas, United States

A method is developed to detect group structures and significant covariates in a high-dimensional panel mixed data sampling (MIDAS) model. The slope coefficients of the model are assumed to be heterogeneous, and group structures exist where the slope coefficients are homogeneous within groups and heterogeneous across groups. A doubly penalized least squares estimator is developed to detect the group structures and the sparsity patterns (i.e., selecting significant/relevant covariates) simultaneously without a-priori knowledge about the number of groups, group structures or sparsity pattern of covariates. It is shown that the proposed penalized estimator enjoys the oracle property and can consistently identify the group structures and the sparsity patterns in large samples. The finite sample performance of the proposed estimator is evaluated through Monte Carlo studies and illustrated with a real dataset in forecasting GDP growth across various countries.

CC424 Room S0.11 COMPUTATIONAL STATISTICS Chair: Andreas Artemiou

C1496: **Polynomial regression on SE(3) with an arbitrary connection**
Presenter: **Johan Aubray**, ENAC, France
Co-authors: Florence Nicol

The problem of estimating the position of a mobile, such as a drone, from noisy position measurements is addressed. The framework of differential geometry is used. More precisely, the trajectory of the mobile is modelled as a Lie group-valued curve, the group in question being the special Euclidean group SE(n), with $n = 2$ or 3. Based on this approach, the goal is to implement this technique in the Lie group SE(3) context. Last year, the idea limited to the Levi-Civita connection was presented. It is now extended to an arbitrary connection. A more general mathematical formulation is established by using differential forms. Applications to simulated data are proposed to illustrate the approach with measurements of the R^2 score. The limitations of such a method and future perspectives are discussed.

C1659: **Combinatorial strategies for greedy regression model selection**
Presenter: **Cristian Gatu**, Alexandru Ioan Cuza University of Iasi, Romania
Co-authors: Georgiana-Elena Pascaru, Petru Sebastian Drumia, Erricos Kontoghiorghes

Greedy step-wise algorithms are an established approach to the regression model selection problem. A main drawback of these methods is the reduced number of submodels that are evaluated in order to select a solution, thus, in general, failing to find the optimum. Three strategies that aim to overcome this issue are investigated: SEL-k, TREE-k and SHUTTLE-k. SEL-k builds on standard forward selection but selects the best k variables at each step instead of the only best one. TREE-k is a method that explores a combinatorial search space, thus increasing the number of submodels that are investigated. Specifically, at each step of the algorithm, a new search branch is considered for each of the best k most significant variables. A branch will terminate either when there are no more significant variables to choose from or when all variables have been considered. SHUTTLE-k aims to obtain good solutions while avoiding a prohibitive computational cost. A list of k submodels is stored. During one iteration, for each of the k submodels, the best k variables are chosen, yielding k^2 submodels (augmentation step). From the resulting k^2 list, the best k submodels are kept for the subsequent iteration (reduction step). Various experiments are conducted on both real and artificially generated datasets in order to assess the three proposed algorithms. The results are presented and discussed.

C1384: **New computational methods for multivariate regression**
Presenter: **Daeyoung Ham**, University of Minnesota, United States
Co-authors: Adam Rothman, Brad Price

New methods are proposed for multivariate regression when the regression coefficient matrix is sparse and the error covariance matrix is dense. It is assumed that the error covariance matrix has equicorrelation across the response variables. Two procedures are proposed; one is based on constant marginal response variance (compound symmetry), and the other is based on general varying marginal response variance. Efficient and fast approximation procedures are also developed for high dimensions. The procedures only require a selection of one tuning parameter, as apposed to the existing joint optimization methods. An approximate Gaussian likelihood minimization is also proposed to guarantee the stability of tuning parameter selection. Through extensive numerical studies, the procedure is showcased to outperform relevant competitors in quality of estimation (prediction) as well as computational cost. With additional comprehensive simulations, the procedures are demonstrated to be robust to model misspecification.

CC471 Room BH (S) 2.01 FINANCIAL AND ECONOMETRIC MODELLING Chair: Michail Karoglou

C0377: **Using regression to enhance an existing closed-form implied volatility formula to widen the range of option moneyness**
Presenter: **Wei-Chung Miao**, National Taiwan University of Science and Technology, Taiwan

In the option pricing literature, there are a number of closed-form formulas for implied volatility that allow for the fast construction of the implied volatility surface. Among others, perhaps the best-performing formula is the version of an existing study developed based on rational approximation. While the formula works nicely for options that are near-the-money, its accuracy deteriorates as the moneyness deviates from 0 (standing for at-the-money), and this greatly limits its applicability. The intention is to develop a closed-form formula that can be applied to options with moneyness away from 0. With this formula accompanied, the original formula can be significantly enhanced and becomes applicable for options with much-widened moneyness. The approach is based on a regression of implied volatility on its explanatory variables. Through the identification of key variables as well as the establishment of the proper functional form, the regression-based closed-form formula is successfully developed with high R-squared and is shown to provide nice accuracy. The numerical results show that when the original formula is enhanced by the regression-based formula, the implied volatility can be accurately calculated with absolute percentage error reduced to lower than 1% for options with moneyness over a wide range.

C1581: **Copula-based trading of cointegrated cryptocurrency pairs**
Presenter: **Masood Tadi**, Prague University of Economics and Business, Czech Republic
Co-authors: Jiri Witzany

A novel pairs trading strategy is introduced based on copulas for cointegrated pairs of cryptocurrencies. To identify the most suitable pairs and generate trading signals formulated from a reference asset for analyzing the mispricing index, linear and nonlinear cointegration tests are employed, and a correlation coefficient measure and different copula families are fit, respectively. The strategy's performance is then evaluated by conducting back-testing for various triggers of opening positions, assessing its returns and risks. The findings indicate that the proposed method outperforms previously examined trading strategies of pairs based on cointegration or copulas in terms of profitability and risk-adjusted returns.

C1284: **The specification of a fractionally integrated factor model**
Presenter: **Dominik Ammon**, University of Regensburg, Germany
Co-authors: Tobias Hartl, Rolf Tschernig

The purpose is to investigate a possible negative effect of unnecessary differencing of nonstationary panel data. A factor model framework is utilized where factors may exhibit fractional integration. Nonstationary factors become more visible when the number of time periods T tend to infinity due to an increasing variance. To mitigate the extreme signal-to-noise ratio, the signals of the fractional integrated factors are adjusted. The analysis demonstrates that the common component can be reliably estimated using principal component analysis as the number of variables N and time periods T tends to infinity. Specifically focusing on approximate dynamic factor models, the standard model selection criteria are established to remain effective for nonstationary and fractionally integrated data, assuming the idiosyncratic component is asymptotically stationary.

Compared to the prior study, wherein the variance of I(1) factors tended towards infinity, the visibility of the factors is decreased by diminishing signals. Building upon the signals, a detrimental effect of differencing is found, as it reduces the influence of nonstationary factors, rendering them practically undetectable.

CC434 Room BH (S) 2.02 PORTFOLIO MANAGEMENT Chair: Lorenzo Mercuri

C0623: Social choice theory and market anomalies for portfolio construction

Presenter: **Alessandra Insana**, University of Messina, Italy
Co-authors: Veronica Guidetti

Social choice theory, traditionally used in voting systems to aggregate preferences, is innovatively applied to portfolio construction, taking into account historical market anomalies associated with specific stock characteristics. In particular, a novel methodology is presented for portfolio management based on the majority judgment ranking approach. Judging stocks using eleven well-known market anomalies, this method allows aggregating the information provided, facilitating the construction of ten portfolios and a long-short strategy. The empirical study conducted on US equities shows that the best portfolio strategy consistently outperforms traditional portfolios sorted by single characteristics. Furthermore, the use of majority judgment demonstrates resilience across various implementation choices, thereby reinforcing its effectiveness. This approach to portfolio management offers a promising avenue for future research and practical applications.

C1341: Multiobjective ESG bond portfolio optimization

Presenter: **Maziar Sahamkhadam**, Linnaeus University, Sweden
Co-authors: Andreas Stephan

The purpose is to develop a copula-based pricing model for forecasting bond returns and to apply it to multi-objective bond portfolio (MOBP) optimization. Evidence of asymmetric tail dependence is provided in the zero-coupon bond yield curve, which is modeled using truncated regular vine copulas. By drawing simulations of the term structure from a copula-based dynamic factor model, step-ahead forecasts are obtained for zero-coupon bonds, which are then used to price both callable and non-callable fixed-coupon bonds. These bond prices are applied to solve convex multi-objective portfolio systems that account for various bond characteristics, including average returns, conditional value-at-risk, distance-to-default, transaction costs, and (option-adjusted) duration and convexity. Utilizing a sample of 879 ESG bonds from Europe over a period from January 2016 to July 2024, the MOBPs based on the proposed approach generate higher returns and Sharpe ratios compared to an equally weighted benchmark portfolio. These portfolios also result in lower tail risk, particularly during the COVID-19 pandemic and the Russo-Ukrainian war. Results further indicate that including issuer-level CO2 emissions as a portfolio attribute leads to portfolios with higher returns but also higher tail risk compared to socially responsible MOBPs based on ESG scores.

C1360: Estimation of risk aversion coefficient for tangency portfolio

Presenter: **Stanislas Muhinyuza**, Linnaeus University, Sweden
Co-authors: Stepan Mazur

The purpose is to investigate the distributional properties of the estimated risk aversion coefficient for the tangency portfolio (TP), assuming that the asset returns follow a multivariate normal distribution and/or a matrix-variate closed skew-normal distribution. Under the normality assumption, a stochastic representation, density function, characteristic function, and high-dimensional distribution are constructed to fully characterize the estimated risk aversion coefficient distribution for both finite and high-dimensional settings. For a matrix-variate closed skew-normal distribution assumption, a stochastic representation is provided that is later used to derive the first and second moments and the high-dimensional distribution of the estimated risk aversion coefficient. Through a simulation study, it is shown that the derived high-dimensional asymptotic distributions provide a good approximations to the exact ones for the finite sample sizes under both distributional assumptions considered.

Monday 16.12.2024 09:10 - 10:50 Parallel Session M – CFECMStatistics2024

CI050 Room Auditorium BAYESIAN METHODS AND APPLICATIONS Chair: Mario Peruggia

C0156: **A Bayesian nonparameteric approach to competing risks**
Presenter: **Antonio Lijoi**, Bocconi University, Italy
Co-authors: Claudio Del Sole, Igor Pruenster

Competing risks arise in several problems in survival analysis. In such a framework, data consists of survival times and an associated cause of death. A new class of priors is displayed for the transition probabilities, which are used in a multi-state modeling approach to competing risks. The proposed specification is obtained through a suitable transformation of a vector of discrete hierarchical random measures. These have been successfully applied in several areas (density estimation, clustering, prediction in species sampling problems, inference with hidden Markov models, etc.), while their uses in survival analysis are very limited. A new strategy that leads to closed-form expressions for marginal, posterior and predictive distributions is illustrated and is based on two main tools: (i) A set of latent variables that naturally arise from the data-generating distribution and can be seen as marks associated with the atoms; (ii) An identity for suitably defined moment measures. The undertaken approach allows for evaluating estimates of the (cumulative) incidence and survival functions and of the so-called prediction curve, which is related to future causes of death.

C0157: **Near-Bayesian methods**
Presenter: **Steven MacEachern**, The Ohio State University, United States

Bayesian methods have proven to be extremely successful when the class of models contains the mechanism that generates the data. Their performance suffers when the data-generating mechanism lies outside the support of the prior distribution; that is when the model is misspecified. Several methods have been proposed to handle model misspecification from a Bayesian perspective. These include methods that are formally Bayesian, those that are arguably Bayesian, and those that have a Bayesian motivation but depart from Bayes. This suggests a need for mechanisms to assess how close a method is to a formal Bayesian method. Several approaches are surveyed, briefly highlighting their strengths and weaknesses and contrasting their performance.

C0158: **Model uncertainty in latent Gaussian models with univariate link function**
Presenter: **Mark Steel**, University of Warwick, United Kingdom
Co-authors: Gregor Zens

A class of latent Gaussian models are considered with a univariate link function (ULLGMs). These are based on standard likelihood specifications (such as Poisson, Binomial, Bernoulli, Erlang, etc.) but incorporate a latent normal linear regression framework on a transformation of a key scalar parameter. Model uncertainty regarding the covariates included in the regression is allowed. The ULLGM class typically accommodates extra dispersion in the data and has clear advantages for deriving theoretical properties and designing computational procedures. Posterior existence is formally characterized under a convenient and popular improper prior, and an efficient Markov chain Monte Carlo algorithm is proposed for Bayesian model averaging in ULLGMs. Simulation results suggest that the framework provides accurate results that are robust to some degree of misspecification. The methodology is successfully applied to measles vaccination coverage data from Ethiopia and to data on bilateral migration flows between OECD countries.

CO282 Room S-2.25 METHODS AND MODELS FOR ENVIRONMENTAL AND ECOLOGICAL DATA I Chair: Domenico Vitale

C0384: **A Bayesian parametric approach to estimate misidentification errors in capture-recapture**
Presenter: **Davide Di Cecco**, Unitelma Sapienza, Italy
Co-authors: Andrea Tancredi

The standard methodologies for capture-recapture and species abundance analysis assume the absence of misidentification errors. However, the presence of such errors is well documented in both contexts. For example, in genotype surveys for animal abundance studies, samples such as fur or faeces are collected in an area and analysed to extract DNA. The occurrence of sequencing errors results in the generation of fictitious genotypes that cannot be linked with other cases and, consequently, leads to erroneous inflation of the number of animals captured once (singletons). In microbial diversity studies, environmental samples, such as soil or water, are analysed to extract DNA or RNA and classify the microbial community into different species. Changes in the species abundance distribution are utilised to assess the impact of factors such as climatic change or the use of chemicals on the health of an ecosystem. Again, sequencing errors lead to the recording of fictitious species, which inflates the number of singletons. A fully Bayesian parametric approach is employed to model spurious singletons as false-negative record linkage errors, and an MCMC algorithm is presented to estimate the posterior distribution of the true number of captures and the number of unsampled units. The MCMC can be specified for different parametric assumptions, and we define exactly the families of distributions for which the algorithm can be adapted.

C0417: **Estimating the causal effect of glyphosate aspersion on coca cultivation in Colombia**
Presenter: **Luisa Scaccia**, University of Macerata, Italy
Co-authors: Perla Rivadeneyra, Luca Salvati

Deforestation and unsustainable practices are posing a severe threat to the invaluable natural treasure represented by the Amazon rainforest. Among other anthropogenic factors, illegal activities driven by organized crime, such as illegal logging, poaching, land grabbing, and coca cultivation, also significantly impact deforestation. In Colombia, aerial fumigation with glyphosate was a central policy to curb coca crops between 1994 and 2015, when this measure was banned following years of opposition and questionings about its effectiveness as well as concerns about its collateral effects on local communities and the environment. Many studies and environmental groups argue that health issues and environmental degradation are a direct consequence of the use of glyphosate. Nevertheless, discussions about bringing back the use of glyphosate have recently resumed, and some groups are determined to reintroduce it. In this context, the aim is to investigate the effectiveness of using fumigation with glyphosate to reduce the spread of coca plantations. A 15-year panel (2000-2015) of the municipalities is used, where coca was detected in 2000, including fixed effects for time and municipalities. To deal with endogeneity issues, aspersion is instrumented with the number of days in which the strength of the wind was below a certain threshold, and aerial fumigation could take place. Estimates provide some evidence that aspersion enhances coca crop extension in most model specifications.

C0606: **A model-based approach for evaluating exposure to extreme events in public health: A case study on the Lazio region**
Presenter: **Edoardo Rosci**, Sapienza University of Rome, Italy
Co-authors: Emiliano Ceccarelli, Giovanna Jona Lasinio, Giada Minelli, Massimo Stafoggia

A new Bayesian model-based approach is introduced to deal with extreme epidemiological events in the Lazio region. Given that extreme temperatures and high pollutant concentrations harm the population, the aim is to develop an innovative method to assess exposure. The proposal leverages the Bayesian approach, potentially providing public decision-makers with tools to allocate health resources based on health risks. It begins by exploring data on extreme temperatures and pollutants (particulate 10 and 2.5, NO2, ozone), considering the geographical features of the region and predictors from the literature. A new modeling proposal of another study is then used, focused on quantile auto-regression models for extreme events. The main output of this stage will be exposure surfaces based on marginal quantiles varying over time. Next, a second-level Bayesian model is implemented to scale to the national level. Among various options, nearest neighbors Gaussian process techniques offer optimal computational and probabilistic properties. The final output will provide tools for assessing exposures in environmental epidemiology, aiding public health applications. Extreme temperature stress will be more relevant among older age groups, while extreme pollution events could be linked to diseases in young people, possibly connecting individual histories to exposure evolution.

C0692: **Analyzing community ecology metabarcoding data using variance partitioning methods**
Presenter: **Massimo Ventrucci**, Department of Statistical Sciences, University of Bologna, Italy
Co-authors: Maria Franco Villoria, Luisa Ferrari, Alex Laini

Community ecology is an exciting field where ecologists and statisticians, in a collective effort, produce tools to investigate the distribution of species over space and the relationship these species have with environmental covariates. Generalized linear mixed models (GLMM) are pervasive in this field as they allow inferring both niche-related processes and dispersal. GLMMs are applied to a specific type of community ecology data called metabarcoding. Metabarcoding is a genetic technique that facilitates species identification through a sequence of operations: preprocessing samples collected in the field, targeting a small region of DNA, and sequencing. This produces large datasets where each sample has an attached label indicating a haplotype, i.e. a variant of a genetic species. One main goal is to separate the variance explained by the environmental covariates (abiotic factors) from the variance explained by the random effects (biotic factors such as dispersal and interaction). Several indicators to separate the contribution of fixed and random effects are proposed in the community ecology literature. These methods are discussed and compared with a novel procedure to perform variance partitioning based on, first, redefining the GLMM so that each model component is represented by a meaningful variance parameter and, second, posterior inference is derived reflecting the explained variance of each term. The method is illustrated with data from a real case study.

CO193 **Room S-1.06** **BIOSTATISTICS AND MACHINE LEARNING: BENCHMARKING, EVALUATION AND BEYOND** **Chair: Roman Hornung**

C0684: **The selection and creation of benchmark data sets for comparison studies: Challenges and solutions**
Presenter: **Silke Szymczak**, University of Luebeck, Germany

Benchmark data sets are crucial to ensure a fair and comprehensive evaluation and comparison of statistical methods. Ideally, a large number of diverse data sets that are representative and relevant to the application area of interest should be used. One approach is to select real-world data sets from publicly available data repositories such as OpenML and UCI. However, a major limitation is the poor documentation of the data sets, which includes missing information on the original source, the main research question, and the interpretation and coding of variables. Specific resources such as TCGA are often used to evaluate approaches for multi-omics analyses, but they focus only on cancer and it is unclear whether the results are transferable to other tissues and diseases. An alternative is to generate synthetic data sets based on predefined statistical models and scenarios. However, it is important that the simulated data are not in favor of any particular statistical method. They should also be as realistic as possible, for example, in terms of correlation structures, noise levels and patterns of missing values. Some solutions from methodological machine learning research for tabular clinical and molecular data are presented and discussed.

C0582: **Evaluating model performance through confidence intervals for the generalization error**
Presenter: **Hannah Schulz-Kuempel**, LMU Munich, Germany

How accurately can a model predict outcomes for new, unseen data? This central question in predictive modeling is addressed by estimating the generalization error (GE), representing the expected loss between model predictions and true outcomes on a new data point. Resampling methods form the crucial basis for the estimation of the GE without actually requiring new data. For resampling-based point estimates to be meaningful, however, precision information is needed in the form of confidence intervals (CIs). Unfortunately, computing a theoretically valid and practically accurate CI for the GE is complicated by the resampling setup. Despite various proposed methods for deriving these CIs, no universal consensus on the best-performing method for different scenarios currently exists. The performance of thirteen model-agnostic methods is benchmarked for deriving CIs for the GE across various supervised learning models and simulation designs, aiming to provide an unbiased assessment of current techniques, establish a foundation for evaluating future methods, and generate hypotheses for further research. Findings form the basis for cautious recommendations for how to compute CIs for the GE and highlight trends and unexpected behaviors, offering insights into the complexities of resampling-based inference. The performance, intricacies of these methods, and their implications are explored for model evaluation in machine learning for biostatistics.

C0917: **On the handling of method failure in comparison studies**
Presenter: **Milena Wuensch**, LMU Munich, Germany
Co-authors: Moritz Herrmann, Elisa Noltenius, Mattia Mohr, Tim Morris, Anne-Laure Boulesteix

Comparison studies in machine learning or classical statistics are intended to compare methods in an evidence-based manner, offering guidance to data analysts to select a suitable method. A common challenge is to handle the failure of some methods to produce a result for some (real or simulated) data sets so that their performances cannot be measured in these instances. Despite an increasing emphasis in recent literature, there is little guidance on proper handling and interpretation, and reporting of the chosen approach is often neglected. The aim is to fill this gap and provide practical guidance for handling method failure in comparison studies. In particular, it is shown that two of the most commonly applied approaches, namely discarding the corresponding data sets (either for all or only the failing methods) and imputing, can lead to misleading method recommendations. It also illustrates how method failure in published comparison studies- such as those in simulation studies and real-data benchmarking across different contexts like regression modelling, statistical testing, and machine learning- may manifest in different ways but is always caused by a complex interplay of two aspects. Based upon this, more adequate recommendations are provided for dealing with method failures that are not based on discarding data or imputation. Finally, the recommendations and the dangers of inadequate handling of method failure are illustrated through two illustrative comparison studies.

C0979: **Multi forests: Variable importance for multi-class outcomes**
Presenter: **Roman Hornung**, University of Munich, Germany
Co-authors: Alexander Hapfelmeier

In prediction tasks with multi-class outcomes, identifying covariates specifically associated with one or more outcome classes can be important. Conventional variable importance measures (Vims) from random forests (Rfs), like permutation and Gini importance, focus on overall predictive performance or node purity without differentiating between the classes. Therefore, they can be expected to fail to distinguish class-associated covariates. A new Vim called multi-class Vim is tailored to identify exclusively class-associated covariates via a novel Rf variant called multi forests (Mufs). The trees in Mufs use both multi-way and binary splitting. The multi-way splits generate child nodes for each class, using a split criterion that evaluates how well these nodes represent their respective classes. This setup forms the basis of the multi-class Vim, which measures the discriminatory ability of the splits performed in the respective covariates with regard to this split criterion. Simulation studies demonstrate that the multi-class Vim specifically ranks class-associated covariates highly, unlike conventional Vims, which also rank other types of covariates highly. Analyses of 121 datasets reveal that Mufs often have slightly lower predictive performance compared to conventional Rfs. This is, however, not a limiting factor given the algorithm's primary goal of calculating the multi-class Vim.

CO225 **Room S-1.27** **STATISTICAL AND MACHINE LEARNING IN ENGINEERING** **Chair: Jan Gertheiss**

C0539: **Monitoring of confounder-adjusted scores using conditional principal component analysis**
Presenter: **Lizzie Neumann**, Helmut Schmidt University, Germany

Co-authors: Philipp Wittenberg, Martin Koehncke, Alexander Mendler, Sylvia Kessler, Jan Gertheiss

In structural health monitoring (SHM), measurements from various sensors are collected and reduced to damage-sensitive features. Diagnostic values for damage detection are then obtained through statistical analysis of these features. The system outputs, i.e., sensor measurements and/or extracted features, however, depend not only on damage but also on confounding factors (environmental or operational variables). These factors affect not only the mean but also the covariance. This is particularly significant because the covariance is often used as an essential building block in damage detection tools. A method is presented for calculating confounder-adjusted scores utilizing conditional principal component analysis, which entails estimating a confounder-adjusted covariance matrix. The technique is applied to monitor real-world data from the Vahrendorfer Stadtweg bridge in Hamburg, Germany.

C0813: **Functional quantile analysis for sensor outputs in structural health monitoring**

Presenter: **Frederike Vogel**, Helmut-Schmidt-University, Hamburg, Germany

Structural health monitoring is a pivotal discipline in determining the condition of a given structure, e.g., a bridge, by gathering and assessing data from sensory systems attached to it. These sensor data can be interpreted as functional. As structural damage can impact the structure's service life, detecting potential damage as quickly as possible is important. A comprehensive analysis of all signals' distributions is essential to achieve this. However, conventional monitoring concepts based on, for instance, functional principal component analysis (FPCA) fall short in accounting for skewness or shifting effects as they merely represent curves as deviations from the mean. In this innovative approach, FPCA is expanded by incorporating a quantile perspective, thereby considering scores at various quantile levels as vital monitoring metrics. Furthermore, the model takes into account confounding effects, specifically the temperature. The method is validated through simulation studies and real-data scenarios.

C1302: **Deriving Gaussian processes for physics-informed structural health monitoring**

Presenter: **Matthew Jones**, University of Sheffield, United Kingdom

Co-authors: Daniel Pitchforth, Elizabeth Cross

In engineering, it is vital that high-value assets such as bridges and aircraft are managed, operated and maintained in both a cost-effective and safe manner. Structural health monitoring has established itself as a promising solution to intelligent asset management, where measurement data obtained from sensors attached to the structure are used to monitor its condition in real time. Machine learning tools such as Gaussian processes are then used to interpret the data and subsequently identify the health state of the structure. One limitation of the use of machine learning tools in structural health monitoring is that training data are required to represent all of the environmental and operational conditions that the structure will experience, in addition to all damage states of interest. In many practical monitoring scenarios, collecting a data set that is extensive is infeasible and limits the effectiveness of data-driven monitoring strategies. The aim is to propose fusing physical knowledge inside Gaussian processes to mitigate against limited training data. By fusing known physics into a data-driven learner, modelling practitioners can combine the expressive power of machine learners with known mechanistic laws, offering enhanced predictive performance where training data is lacking, in addition to improving model generalization. The method will be demonstrated in a real-life case study.

C1344: **Quantifying the value of domain knowledge in physics-informed machine learning**

Presenter: **Aidan Hughes**, The University of Sheffield, United Kingdom

Co-authors: Elizabeth Cross, Keith Worden, Sam Gibson, Timothy Rogers, Matthew Jones

Predictive models are used to support decision-making in a variety of tasks within the field of engineering, including design, control, and maintenance planning. Recent research into physics-informed machine learning (or grey-box modelling) has sought to overcome the limitations of models based solely on physics or data. Physics-informed machine learners possess enhanced predictive capabilities, retaining the flexibility of data-based models while also being able to leverage the domain knowledge encoded in physical equations and constraints for regions of the input space where data are scarce. Examining the different modelling approaches from a decision-theoretic perspective is desirable to highlight the advantages of physics-informed machine learning models to end-users and to accelerate the adoption of such models into engineering decision processes. To this end, an approach is presented for quantifying the value of information associated with the domain knowledge present in physics-informed machine learning models. The value of information is a concept from decision theory that captures the difference in the maximum (expected) utilities achieved for a given decision process when solved with and without said information. The approach is demonstrated using a simulated decision problem framed around a numerical case study pertaining to the operation and maintenance of a structure.

CO086 Room K0.18 IMPACTFUL SPATIO-TEMPORAL STATISTICAL APPLICATIONS Chair: Peter Craigmile

C0192: **A dynamic spatiotemporal stochastic volatility model with an application to environmental risks**

Presenter: **Philipp Otto**, University of Glasgow, United Kingdom

Co-authors: Osman Dogan, Suleyman Taspinar

A dynamic spatiotemporal stochastic volatility (SV) model is introduced, incorporating explicit terms accounting for spatial, temporal, and spatiotemporal spillover effects. Alongside these features, the model encompasses time-invariant site-specific factors, allowing for differentiation in volatility levels across locations. The statistical properties of an outcome variable within this model framework are examined, revealing the induction of spatial dependence in the outcome variable. Additionally, a Bayesian estimation procedure employing the Markov chain Monte Carlo (MCMC) approach, complemented by a suitable data transformation, is presented. Simulation experiments are conducted to assess the performance of the proposed Bayesian estimator. Subsequently, the model is applied in the domain of environmental risk modeling, addressing the scarcity of empirical studies in this field. The significance of climate variation studies is emphasized, illustrated by an analysis of local air quality in Northern Italy during 2021, which underscores pronounced spatial and temporal clusters and increased uncertainties/risks during the winter season compared to the summer season.

C0569: **Spatial extreme value modelling via a geometric approach**

Presenter: **Lydia Kakampakou**, Lancaster University, United Kingdom

Co-authors: Jenny Wadsworth

Recent developments in extreme value statistics research have established the geometric approach as a powerful modelling tool for multivariate extremes. Such an approach has not yet been studied within a spatial framework. A novel approach is developed for the modelling of spatial extremes based on spatial gauge functions. Gauges are investigated for several known spatial models, and inference is performed on these by extending methodology from the modelling framework of Wadsworth and Campbell. Simulation studies suggest promising results for parameter estimation and extrapolation via simulation from the model to estimate extreme set probabilities. If time permits, an application to a space weather dataset will be illustrated.

C0592: **Quantifying causal relationships from climate observations using spatiotemporal stochastic interventions**

Presenter: **Samuel Baugh**, Pennsylvania State University, United States

Physical dynamics indicate a causal relationship between greenhouse gas concentrations and a changing climate, but estimating the exact magnitude of expected warming is a notoriously difficult problem. Inferring the effect from observations alone can be framed as an observational causal inference problem; however, in addition to the classical challenge of accounting for unobserved counterfactual outcomes, causal inference in the climate system is particularly difficult due to the fact that there is essentially a single observation. The aim is to propose addressing these challenges by extending the stochastic intervention framework to continuous, physically informed spatiotemporal processes. By specifying the distributional form of the stochastic intervention, this method can consistently estimate the spatially-varying causal effect from observations. As these distributions cannot be estimated from observations alone without unrealistically strong assumptions, a framework is proposed for additionally incorporating prior information from climate model simulations to constrain the estimation. The robustness of the resulting method is assessed through sensitivity analyses and validation studies using withheld climate model data.

C0524: **Gaussian process models for pollution in rivers**
Presenter: **Theresa Smith**, University of Bath, United Kingdom
Co-authors: Marno Basson, Tobi Louw

The impact of human activity on the quality of surface waters, including rivers, has recently garnered considerable attention in the media. Statistical models to characterize the spatiotemporal distribution of biological and chemical indicators in a river network must accommodate several features not seen in typical spatial modelling applications, including censoring of measured concentrations and complicated representations of distance. A Bayesian approach is presented that addresses these two challenges within the framework of the spatial stream network model of another study applied to water quality monitoring data in Stellenbosch, South Africa.

CO084 Room K0.19 NEW APPROACHES AND APPLICATIONS OF SPATIAL STATISTICS Chair: Nicoletta D Angelo

C0363: **A roughness penalty approach for time-evolving occurrences on planar and curved regions**
Presenter: **Blerta Begu**, University College Dublin, Italy

The purpose is to address space-time point processes and analyze their continuous evolution across both spatial and temporal dimensions. An innovative nonparametric methodology is introduced to estimate the unknown space-time density of point patterns or, equivalently, the intensity of an inhomogeneous space-time Poisson point process. The approach combines maximum likelihood estimation with roughness penalties, leveraging differential operators across spatial and temporal domains. Key theoretical properties of the estimator are first established, including consistency. Subsequently, an efficient and flexible estimation procedure is developed, utilizing advanced numerical and computational techniques. By employing finite elements for spatial discretization and B-splines for temporal discretization, the method can effectively model complex, multi-modal, and strongly anisotropic spatiotemporal point patterns. These patterns can be observed over planar or curved domains with intricate geometries, such as coastal regions with complicated shorelines or curved regions with complex orography. Beyond estimation, the method includes tools for appropriate uncertainty quantification. The proposed method is validated through simulation studies and applications to real-world data, demonstrating significant advantages over existing state-of-the-art approaches.

C0556: **Survival modelling of smartphone trigger data for earthquake parameter estimation in early warning**
Presenter: **Luca Aiello**, University of Milano Bicocca, Italy
Co-authors: Lucia Paci, Raffaele Argiento, Francesco Finazzi

Crowdsourced smartphone-based earthquake early warning systems recently emerged as reliable alternatives to the more expensive solutions based on scientific-grade instruments. For instance, during the 2023 Turkish-Syrian deadly event, the system implemented by the Earthquake Network citizen science initiative provided a forewarning of up to 25 seconds. A statistical methodology is developed based on a survival mixture cure model, which provides full Bayesian inference on epicenter, depth and origin time, and an efficient tempering MCMC algorithm is designed to address the multi-modality of the posterior distribution. The methodology is applied to data collected by the Earthquake Network, including the 2023 Turkish-Syrian and 2019 Ridgecrest events.

C0784: **Estimating rare species distribution with opportunistic data: The case of the white shark in the Mediterranean Sea**
Presenter: **Greta Panunzi**, Sapienza University of Rome & University of Salento, Italy

Preserving apex predators in the ocean is extremely important. The lack of comprehensive abundance and distribution data often hinders our understanding of the population status of many endangered species. Occurrence records are usually limited and opportunistic, and fieldwork to gather more data is expensive and frequently not successful. Therefore, it is crucial to combine different sources of data to create models of species distribution that can help guide sampling and conservation efforts. The white shark is a rare but persistent inhabitant of the Mediterranean Sea, classified as critically endangered by the IUCN. Despite this, population abundance, distribution patterns, and habitat use remain poorly understood. Available occurrence records from 1985 to 2021 from diverse sources are utilized to construct a spatial log-Gaussian Cox process, incorporating data-source-specific detection functions and thinning and accounting for physical barriers. The model estimates white shark presence intensity and uncertainty using a Bayesian approach with integrated nested Laplace approximation (INLA). For the first time, species occurrence hotspots and landscapes of relative abundance (continuous measures of animal density in space) are projected throughout the Mediterranean Sea. This approach can be applied to other rare species for which presence-only data from different sources are available, enhancing conservation strategies.

C0944: **SPDE-Forest: An hybrid approach for modeling geostatistical data**
Presenter: **Luca Patelli**, University of Pavia, Italy
Co-authors: Michela Cameletti, Mario Figueira Pereira

The aim is to propose SPDE-Forest, a hybrid two-stage approach for modeling geostatistical data and performing spatial prediction. The proposed strategy combines the random forest algorithm and the stochastic partial differential equations (SPDE) approach implemented through the INLA algorithm. SPDE-Forest is able to model complex non-linear relationships between the response variable and the predictors, thanks to the random forest stage, while accounting for the residual spatial correlation by means of the SPDE part. The out-of-sample predictive performance of SPDE-Forest is assessed through a simulation study and compared to that of existing competitors.

CO383 Room K0.20 DATA DEPTH FOR COMPLEX DATA TYPES AND APPLICATIONS Chair: Pavlo Mozharovskyi

C0373: **Data depth for probability measures**
Presenter: **Myriam Vimond**, ENSAI, France
Co-authors: Pavlo Mozharovskyi, Pierre Lafaye de Micheaux

Statistical data depth measures the centrality of a given point in space with respect to a finite sample or with respect to a probability measure in that space. Over the last few decades, this seminal idea of data depth has evolved into a powerful tool that has proven useful in various fields of science. Recently, the notion of data depth was extended to unparametrized curves. A notion of data depth is proposed, which is suitable for data represented as probability measures. Applications with finite finite point processes are considered, with distributions of random closed sets, or with models of germ grain coverage. Depending on the geometry of the data, adaptations of this depth are investigated, for example, by introducing a weight. It is shown that the depth satisfies the theoretical requirements of general depth functions that are meaningful for applications.

C0701: **Halfspace depth as a classification loss: A machine learning viewpoint on statistical data depth**
Presenter: **Arturo Castellanos**, Telecom Paris, France
Co-authors: Pavlo Mozharovskyi, Hicham Janati

Data depth is a score function that quantifies how deep a point is inside a distribution (or a data set) and has applications in multivariate analysis, anomaly detection, classification, and statistical testing, to name a few. Historically, the very first notion of data depth has been the halfspace depth introduced by John W. Tukey, which generalises the notion of quantile to the multivariate setting. Taking a different angle from the quantile point of view, it is shown that halfspace depth can also be regarded as the minimum loss of a set of classifiers for a specific labelling of the observations.

A natural extension proposed is to change to different sets of classifiers, well-known in the machine learning literature, such as support vector machines or neural networks. Properties such as statistical convergence and speed of the optimization programs are naturally inherited from the literature on those classifiers. How theory can help pick the hyperparameters to get the most sensible results is discussed, with supportive simulations and experiments on data.

C0747: **Local depth functions and clustering**
Presenter: **Giacomo Francisci**, University of Ulm, Germany
Co-authors: Claudio Agostinelli, Alicia Nieto-Reyes, Anand Vidyashankar

Local depth functions are a generalization of depth functions and are used to capture local features of multivariate distributions. When the distribution is absolutely continuous, rescaled local depth functions converge uniformly to the underlying density. Under appropriate regularity conditions, their derivatives also converge. Using these results and a gradient system analysis, it is developed a clustering algorithm based on identification of (i) the modes and (ii) the basis of attractions of the modes via the gradient system. The algorithm is consistent in the sense that the probability distance between true and empirical clusters converges to zero as the sample size diverges to infinity. To show this, it is established a Bernstein-type inequality for deviations between the centered and rescaled local depth functions. Finally, the finite sample performance of the algorithm is investigated via Monte Carlo simulations.

C0937: **MDS-based depth for mixed-type data applied to the assessment of biological age**
Presenter: **Ignacio Cascos**, Universidad Carlos III de Madrid, Spain
Co-authors: Aurea Grane Chavez, Jingye Qian

In a mixed-type dataset, a new procedure to evaluate the centrality of an observation is introduced. This method is then used to assess the biological age of an individual, which is derived from biomarkers, medical conditions, life habits, and sociodemographic variables. These records are mixed-type, encompassing both numerical and categorical variables some of which are nonbinary. To measure the centrality of an observation within such a dataset, Gower's distance is employed between each pair of objects to build an Euclidean representation of the dataset through multidimensional scaling. Finally, classical multivariate data depth notions in such space are used. Ultimately, an individual's biological age is evaluated by finding the age that positions its records as centrally as possible among a sample of similar-aged individuals, keeping all other features constant.

CO203 Room K0.50 DESIGN OF EXPERIMENTS AND APPLICATIONS Chair: Stella Stylianou

C1606: **Sliced designs for computer experiments using sequences with zero autocorrelation function**
Presenter: **Omar Alhelali**, RMIT university, Australia

Computer experiments have become increasingly important because they save time and money compared to physical experiments. A special type of experimental design, called sliced designs, is discussed, which expands on the idea of sliced Latin hypercube designs. The focus is on developing unique sliced orthogonal designs for computer experiments that can be used in first- and second-order models. Each slice forms a smaller, orthogonal design that's well-suited for computer simulations. The construction method proposed relies on using sequences like T-sequences, Golay sequences, and disjoint amicable sequences, which have zero autocorrelation functions. This method allows creating the first infinite families of these designs.

C1555: **Supersaturated design-based statistical methods for variable selection in high-dimensional observational data**
Presenter: **Tharkeshi Dharmaratne**, RMIT University, Australia
Co-authors: Alysha De Livera, Stelios Georgiou, Stella Stylianou

In experimental studies, supersaturated designs (SSDs)-based statistical methods are commonly used to screen relevant factors when the number of factors exceeds the run size. Based on simulation studies, several of these SSD-based statistical methods have been shown to perform well in experimental settings. It motivated the exploration of using these SSD methods on high-dimensional observational data for variable selection. Variable selection is a widely used approach for selecting variables of a statistical model in observational studies, which has often been criticized. Therefore, initially reviewed the latest recommendations and methods that are developed for variable selection in observational studies. The performance of the SSD-based statistical methods is then evaluated using both simulated and real-life data, followed by a comparison of their performance with the existing approaches.

C1611: **Enhancing response surface methodology with Latin hypercube sampling techniques**
Presenter: **Despina Athanasaki**, RMIT, Australia

A novel method is presented for incorporating Latin hypercube sampling (LHS) into composite designs within the design of experiments (DOE) for response surface methodology (RSM). Composite designs are known for their flexibility and efficiency in probing complex relationships between input variables and output responses. By integrating LHS, the exploration of multidimensional parameter spaces is significantly enhanced, ensuring a more diverse and representative sampling scheme. Practical examples and comprehensive analysis demonstrate how LHS can innovate composite designs in RSM applications. Results highlight LHSs potential to optimize processes, improve product quality, and reduce costs across various fields, offering valuable insights for both practitioners and researchers in experimental design and optimization.

C1664: **Computational construction of sequential efficient designs for the second order model**
Presenter: **Norah Alshammari**, student, Australia
Co-authors: Stelios Georgiou, Stella Stylianou

Sequential experiment designs optimize data collection by enabling efficient decisions on whether to continue or stop testing, minimizing resource usage and accelerating goals. Sequential Latin hypercube designs (SLHDs) progressively add design points, reducing computational demands. Unlike traditional model-free LHDs, this approach generates optimal designs for specified models, focusing on the second-order model used in response surface methodology to improve A-efficiency. Challenges in designing efficient high-dimensional experiments are addressed by relaxing the condition of no-point replication in equally spaced intervals. This relaxation maintains space coverage while allowing more flexibility for model-specific efficiency. Sobol sequences are used to iteratively select points that maximize the A-criterion for the second-order model. Results show superior performance compared to methods that minimize inner-point distances, offering practical guidance for selecting optimal experimental designs.

CO260 Room K2.31 (Nash Lec. Theatre) ADVANCES IN CAUSAL INFERENCE Chair: Stathis Gennatas

C1023: **Evaluating and utilizing surrogate outcomes in covariate-adjusted response-adaptive designs**
Presenter: **Wenxin Zhang**, UC Berkeley, United States

The intersection of surrogate outcomes and adaptive designs in statistical research is explored. Current surrogate evaluation methods do not directly account for the benefits of using surrogate outcomes to adapt randomization probabilities in trials, which aim to address treatment effect heterogeneity. Surrogate outcomes can minimize participant regret by enabling rapid adaptation of randomization probabilities, especially when early detection of heterogeneous treatment effects is possible. A novel approach is introduced for surrogate evaluation in sequential adaptive designs, and a new covariate-adjusted response-adaptive (CARA) design is proposed using an online superlearner. This approach adaptively chooses surrogate outcomes to update treatment randomization probabilities. A targeted maximum likelihood estimation (TMLE) estimator is presented to address data dependency challenges, achieving asymptotic normality under reasonable assumptions without relying on parametric models. Simulations demonstrate the robust performance of the adaptive design. The framework not only provides a method to quantify the benefits of surrogate outcomes but also offers an easily generalizable tool for evaluating various adaptive designs and making inferences, providing insights into alternative choices of designs.

C1597: **Evaluating and improving real-world evidence with targeted learning**
Presenter: **Rachael Phillips**, University of California Berkeley, United States

Society is drowning in data and the current practice of learning from data is to apply traditional statistical methods that are overly simplistic, arbitrarily chosen, and subject to manipulation. Nonetheless, these methods inform policy and science, affecting our sense of reality and judgements. The aim is to expose this practice and present a solution, a principled and reproducible approach, termed targeted learning, for generating actionable and truthful information from complex, real-world data. This approach unifies causal inference, machine learning and deep statistical theory to answer causal questions with statistical confidence.

C0310: **Treatment effects in staggered adoption designs with non-parallel trends**
Presenter: **Emmanuel Tsyawo**, Universite Mohammed VI Polytechnique, Morocco
Co-authors: Brantly Callaway

The purpose is to consider identifying and estimating causal effect parameters in a staggered treatment adoption setting - that is, where a researcher has access to panel data and treatment timing varies across units. The case is considered where untreated potential outcomes may follow non-parallel trends over time across groups. This implies that the identifying assumptions of leading approaches, such as difference-in-differences, do not hold. The main focus is on the case where untreated potential outcomes are generated by an interactive fixed effects model and show that variation in treatment timing provides additional moment conditions that can be used to recover a large class of target causal effect parameters. The approach exploits the variation in treatment timing without requiring either (i) A large number of time periods or (ii) Any extra exclusion restrictions. This is in contrast to essentially all of the literature on interactive fixed effects models, which requires at least one of these extra conditions. Rather, the approach directly applies in settings where there is variation in treatment timing. Although the main focus is on a model with interactive fixed effects, the idea of using variation in treatment timing to recover causal effect parameters is quite general and could be adapted to other settings with non-parallel trends across groups, such as dynamic panel data models.

C0393: **Inference in regression with latent clusters: A penalty-free approach**
Presenter: **Abdul-Nasah Soale**, Case Western Reserve University, United States

Model misspecification is a common problem in estimating treatment effects in samples from heterogeneous populations. The problem of estimating treatment effects in regression involving latent clusters due to samples coming from populations with different means and latent clusters induced by the inclusion of ill-defined categorical predictors in the model is presented. A two-step procedure for model-based subgroup analysis involving k-means clustering and least squares regression is proposed for estimating proportional and cluster-varying treatment effects. The proposed method also provides a graphical check for endogeneity between the latent clusters and the observed predictors via sufficient summary plots. The performance of the method on synthetic and two real data applications on medical and heating costs are included. The theoretical justifications are also provided.

C1084: **Assessing causal effects of radiation toxicity on overall survival**
Presenter: **Gilmer Valdes**, UCSF, United States

The causal effects of radiation-induced toxicities are investigated on overall survival (OS) in stage III non-small cell lung cancer (NSCLC) patients undergoing proton radiation therapy. Utilizing real-world data (RWD) from the Registry Study for Radiation Therapy Outcomes and advanced causal inference methodologies supported by the US Food and Drug Administration's Real-World Evidence (RWE) Program, the impact of pneumonitis/dyspnea and esophagitis/dysphagia is estimated on OS using targeted minimum loss-based estimation (TMLE). Results indicate a potential time-dependent relationship between pneumonitis/dyspnea and reduced OS, highlighting the importance of personalized treatment strategies that balance treatment intensity with toxicity management. Furthermore, the approach demonstrates the value of targeted learning and its systematic causal roadmap in generating reliable RWE. This sets a precedent for applying these techniques in radiation oncology to support clinical and regulatory decision-making and ultimately improve patient care.

CO136 Room K2.40 STATISTICAL SEQUENTIAL METHODS FOR DECISION-MAKING PROBLEMS Chair: Matteo Borrotti

C0542: **Optimal subsampling for hierarchical data**
Presenter: **Songqiao Han**, Kings College London, United Kingdom
Co-authors: Kalliopi Mylona, Steven Gilmour

Hierarchical data analysis is an important topic in big data research. However, the computational costs associated with parameter estimation and model fitting in large datasets are very high, making efficient subsampling techniques necessary. An optimal subsampling method is introduced specifically designed for hierarchical data, which fully considers the connections between and within different levels. This method enables optimal subsampling across various scenarios. In addition, several examples of industrial data applications are given.

C0488: **Enhancing data efficiency in online deep reinforcement learning under partial observability**
Presenter: **Valentina Zangirolami**, University of Milano-Bicocca, Italy

Model-free (MF) methods have been widespread in online deep reinforcement learning (DRL) literature due to their good asymptotic performance and reliance only on policy estimation. DRL effectively addresses complex real-world scenarios with high dimensional states, leveraging scalable neural networks. Such scenarios can also be affected by the partial observability of states and can be described by partially observable Markov decision processes (POMDPs). MF methods typically require many interactions, resulting in data efficiency issues. The aim is to propose a novel model-based (MB) DRL method called deep recurrent Dyna-Q, which adapts the existing deep Dyna-Q framework to partial observability. MB-DRL introduces the concept of planning, which consists of interacting with the learned POMDP dynamics. Essentially, Dyna methods combine MF and MB-DRL, where the value function is updated employing both the real and simulated experience from the learned dynamics, thus leveraging the asymptotic performance of MF while improving data efficiency through MB. Variational recurrent neural networks are used as a model for estimating the conditional density of observations involving additional stochasticity in hidden states. Experiments are conducted using this novel framework for self-driving cars with different sample update methods, providing a comprehensive statistical analysis and benchmarking against state-of-the-art methods.

C0615: **Sequential knockoffs for variable selection in reinforcement learning**
Presenter: **Jin Zhu**, London School of Economics and Political Science, United Kingdom

In real-world applications of reinforcement learning, it is often challenging to obtain a state representation that is parsimonious and satisfies the Markov property without prior knowledge. Consequently, it is common practice to construct a state larger than necessary, e.g., by concatenating measurements over contiguous time points. However, needlessly increasing the dimension of the state can slow learning and obfuscate the learned policy. The notion of a minimal sufficient state is introduced in a Markov decision process (MDP) as the smallest subvector of the original state under which the process remains an MDP and shares the same reward function as the original process. A novel SEEK algorithm is proposed that estimates the minimal sufficient state in a system with high-dimensional complex nonlinear dynamics. In large samples, the proposed method achieves selection consistency. As the method is agnostic to the reinforcement learning algorithm being applied, it benefits downstream tasks such as policy learning. Empirical experiments verify theoretical results and show the proposed approach outperforms several competing methods regarding variable selection accuracy and regret.

C0855: **Optimal design for A/B testing in time series experiments**
Presenter: **Chengchun Shi**, LSE, United Kingdom

Time series experiments, in which experimental units receive a sequence of treatments over time, are frequently employed in many technological companies to evaluate the performance of a newly developed policy, product, or treatment relative to a baseline control. Many existing A/B testing solutions assume a fully observable experimental environment that satisfies the Markov condition, which often does not hold in practice. The optimal design for A/B testing is studied in partially observable environments. A controlled (vector) autoregressive moving average model is introduced to capture partial observability. A small signal asymptotic framework is introduced to simplify the analysis of asymptotic mean squared errors of average treatment effect estimators under various designs. Two algorithms are developed to estimate the optimal design: one utilizing constrained optimization and the other employing reinforcement learning. The superior performance of the designs is demonstrated using a dispatch simulator and two real datasets from a ride-sharing company.

CO173 Room K2.41 SURVIVAL AND LONGITUDINAL DATA ANALYSIS Chair: Din Chen

C0717: **Bayesian mixture of accelerated-failure-time experts model**
Presenter: **Elham Mirfarah**, University of St Andrews, United Kingdom

Exploring the relationship between survival time and the covariates of interest is a challenging topic in biomedical studies. The classical accelerated failure time (AFT) model is often more flexible, powerful, and interpretable than the Cox proportional hazards model if the underlying assumptions (distributional and homoscedasticity) are met. However, real-world data often exhibit heteroscedasticity, which compromises the robustness of the classical AFT model. A parametric approach to addressing non-homogeneous datasets is the mixture-of-experts (MoE) models. The MoE is an extension of the finite mixture model wherein the mixing proportion varies for each observation. A Bayesian analysis is presented for censored survival time data, employing a broad class of distributions (scale mixture of normal) for the error term in the AFT model. A weakly informative prior structure is proposed for the parameters, and the corresponding posterior distributions are demonstrated to be proper. By leveraging the Ultimate Polya-Gamma data-augmentation method, gating parameters are efficiently sampled, and cluster memberships are allocated for data subgroups. The effectiveness of the proposal is illustrated through synthetic studies and a real data example.

C0713: **Robust Bayesian inference for accelerated failure time models with skewed and heavy-tailed survival data**
Presenter: **Mehrdad Naderi**, Northumbria University Newcastle, United Kingdom

Censored data analysis has been commonly used in clinical studies, where researchers often encounter limitations in measuring instruments and/or experimental design. Despite the growing adoption of survival statistical methods for analyzing censored data, the accurate specification of these models requires great care, especially in dealing with atypical observations displaying non-normal features. In such cases, models based on normality may fail to capture the underlying pattern adequately. Accelerated failure time modeling is proposed for survival-censored data in which the random errors conform to the normal mean-variance mixture (NMVM) distribution. The NMVM represents a diverse class of distributions offering enhanced flexibility in analyzing skewed and heavy-tailed data. The focus lies on Bayesian inference for model parameters, employing an efficient Markov chain Monte Carlo (MCMC) algorithm. A series of simulation studies are conducted to assess the effectiveness and robustness of the proposed methods in various scenarios of skew data with heavy tails. The methodology is then examined by assessing the relationship between covariates and survival time on real-world data. The results show that models with a skew and/or fat-tail distribution not only exhibit a superior fit to the data but also provide robust inference compared to the normally-based model.

C0168: **Estimate COVID-19 vaccine efficacy with time-to-infection outcome**
Presenter: **Din Chen**, University of Pretoria, South Africa

The COVID-19 pandemic has caused significant morbidity and mortality, as well as social and economic disruption worldwide. In order to reduce these effects, a global effort to develop effective vaccines against the COVID-19 virus has produced various options with the effectiveness assessed on the rate of infection between vaccinated and unvaccinated groups, which has been used for important policy decision-making on vaccination effectiveness ever since. However, the rate of infection is an over-simplified index in assessing the vaccination effectiveness overall, which should be strengthened to address the duration of protection with time-to-infection effect. The fundamental challenge in estimating the vaccination effect over time is that the time-to-infection for the unvaccinated group is unknown due to nonexistent vaccination time. The purpose is to discuss the biostatistical methodological development to fill this knowledge gap and propose a Weibull regression model. This model treats the nonexistent vaccination time for the unvaccinated group as nuisance parameters and estimates the vaccination effectiveness along with these nuisance parameters. The performance of the proposed approach and its properties are empirically investigated through a simulation study, and its applicability is illustrated using a real-data example from the Arizona State University COVID-19 serological prevalence data.

C0259: **Dynamic prediction with numerous longitudinal covariates made easy: The R package pencal**
Presenter: **Mirko Signorelli**, Leiden University, Mathematical Institute, Netherlands

To make informed decisions, clinicians and patients rely on accurate predictions of the probability of experiencing adverse events (such as dementia, cancer, or death) over time. Dynamic prediction models make it possible to update the probability of experiencing an event as more longitudinal data is collected. Traditional approaches to dynamic prediction include joint modelling, which is computationally unfeasible with numerous longitudinal predictors, and landmarking, which only uses data from the last available observation. Penalized regression calibration (PRC) is introduced, a dynamic prediction method that is capable of handling numerous longitudinal covariates as predictors of survival. After illustrating the statistical methodology that PRC is based on, how the R package pencal makes it easy to estimate PRC is shown, and the predicted survival probabilities are computed and updated to validate the predicted performance of the fitted model. The results of a systematic comparison of the predictive performance of PRC and alternative modelling approaches are presented using several real-world datasets that differ in terms of survival outcome, sample size, number of longitudinal covariates, and length of the follow-up.

CO154 Room S0.11 RECENT DEVELOPMENT OF STATISTICAL METHODS FOR HANDLING COMPLEX DATA Chair: Sollie Millard

C0193: **Thresholding-based robust estimation for generalized mixture models**
Presenter: **Zhen Zeng**, Nanjing University of Finance and Economics, China

Finite mixture regression models are versatile tools for analyzing mixed regression relationships within clustered and heterogeneous populations. However, the classical normal mixture model often falls short when dealing with nonnormal or nonlinear regression data, especially in the presence of severe outliers. To address this, a novel generalized robust mixture regression procedure is introduced within the finite mixture regression framework. This procedure features sparse, scale dependent mean shift parameters, facilitating outlier detection and ensuring robust parameter estimation. The approach incorporates three key innovations:(1) A penalized likelihood approach using a combination of L0 (zero norm) and L2 (ridge) regularization to induce sparsity among mean shift parameters. (2) A close connection to the method of trimming, including explicit outlyingness parameters for all samples, simplifies computation, aids theoretical analysis, and eliminates the need for parameter tuning. (3) High scalability, allowing the implementation to handle nonnormal or nonlinear regression data. A threshold-based generalized expectation maximization algorithm has been developed to ensure stable and efficient computation. Simulation studies and real-world data applications demonstrate the effectiveness of this robust estimation procedure.

C0285: **Semiparametric inference on inequality measures with nonignorable nonresponse**
Presenter: **Chunlin Wang**, Xiamen University, China

Measuring inequality of economic variables, such as income, is vital in economics, social science and statistics. Reliable estimation and inference of inequality measures can provide insights and have crucial implications in policy-making procedures. However, income survey data inevitably suffer from nonignorable nonresponse, in the sense that the response probabilities depend on the missing income values. This creates challenges for the estimation of inequality measures, in particular, the model identifiability and selection bias issues. To address these issues, we exploit the commonly available callback data in income surveys and propose a semiparametric modeling strategy. We develop a semi-parametric full-likelihood approach for making inference on inequality measures with nonignorable nonresponse. We establish large-sample properties of the proposed estimators of the inequality measures, including the quantile, Gini index, Theil index and the generalized entropy class. Additionally, we devise a stable expectation-maximization algorithm for efficient computation. The simulation results and a real income data example demonstrate that proposed method corrects the bias of the estimated inequality measures and leads to reliable inference results.

C1617: **Predictive modelling with ensembles of projected nearest neighbors**
Presenter: **David Hofmeyr**, Lancaster University, United Kingdom

Nearest neighbors (NN) based estimators are one of the most popular among non-parametric smoothing techniques, largely for their simplicity and (relative) computational efficiency. However, their successful use within ensemble models has been overshadowed by the far more popular decision tree (DT) based models; such as random forests (RF) and gradient boosting (GB) models, which use DTs as base learners. Where DTs are advantaged over (most) other non-parametric smoothers is in how they adaptively determine smoothing neighbourhoods, which are optimised with reference to the problem objective. This adaptiveness naturally increases the complexity of the resulting model, increasing its variance, which is why their greatest successes have been in the context of the ensemble models mentioned previously. Inspired by this, the utility of using projection methods are investigated, which are problem-specific, to allow the neighborhood search in NNs to similarly be adaptive to the problem objective. The resulting models are simple to implement and computationally efficient, as well as being versatile and widely applicable. In addition, their accuracy in prediction tasks, such as classification and regression, is competitive with the DT alternatives.

C0815: **Robust feature and component selection in mixture regression**
Presenter: **Sollie Millard**, University of Pretoria, South Africa
Co-authors: Salomi Millard, Frans Kanfer

A penalized likelihood approach is proposed for model selection in mixtures of generalized linear models. Penalties are imposed on both the mixing proportions and regression coefficients. This enables both component and feature selection. A self-paced learning approach is also proposed to mitigate the impact of outliers. An EM-type algorithm is proposed to maximize the penalized log-likelihood function. The properties of the estimation algorithm are demonstrated using a simulation approach. The proposed estimation approach is also illustrated using real data.

CO066 Room S0.12 RECENT DEVELOPMENTS IN CLUSTERING FOR COMPLEX DATA STRUCTURE Chair: Maria Brigida Ferraro

C0672: **COVID-19 in Italy: Contrasting pre-vaccine epidemic waves through functional data clustering**
Presenter: **Lorenzo Testa**, Carnegie Mellon University, United States
Co-authors: Tobia Boschi, Jacopo Di Iorio, Marzia Cremona, Francesca Chiaromonte

Data from 107 Italian provinces is used to characterize and compare mortality patterns during the first two COVID-19 waves before vaccines were introduced. Using functional data analysis clustering techniques, differences between the two waves are documented, focusing on their magnitude and variability. Specifically, while both waves were characterized by a co-occurrence of 'exponential' and 'mild' mortality patterns, the first had higher and more concentrated mortality peaks, while the second spread much more broadly and asynchronously through the country. Notwithstanding limitations in the accuracy and reliability of publicly available data, these patterns are also associated with mobility, the timing of government restrictions, and socio-demographic, infrastructural, and environmental covariates. Evidence of a significant positive association between local mobility and mortality is found in both epidemic waves, and the effectiveness of timely restrictions is corroborated in curbing mortality. The techniques described could capture additional and potentially sharper signals if applied to richer data.

C0758: **Mixture-based clustering with covariates for ordinal responses**
Presenter: **Marta Nai Ruscone**, Universita degli Studi di Genova, Italy
Co-authors: Daniel Fernandez, Kemmawadee Preedalikit, Louise McMillan, Ivy Liu, Roy Costilla

Existing methods can perform likelihood-based clustering on a multivariate data matrix of ordinal responses, using finite mixtures to cluster the rows and columns of the matrix. Those models can incorporate the main effects of individual rows or columns and the cluster effects to model the matrix of responses. However, many real-world applications also include available covariates. Those mixture-based models are extended to include covariates and to examine what effect this has on the resulting clustering structures. The focus is on clustering the rows of the data matrix using the proportional odds version of the cumulative logit model for ordinal data. The models fit using the expectation-maximization (EM) algorithm, and their performance was assessed using a comprehensive simulation study. Finally, an application of the proposed models is also illustrated in the well-known arthritis clinical trial data set.

C0807: **Clustering longitudinal mixed data**
Presenter: **Francesco Amato**, University Lyon II, France
Co-authors: Julien Jacques

A model-based clustering algorithm is presented to cluster longitudinal mixed data. Assuming that the non-continuous variables are the discretization of underlying latent continuous variables, the model relies on a mixture of matrix-variate normal distributions, accounting simultaneously for within- and between-time dependence structures. The model is thus able to concurrently handle the heterogeneity, the association among the responses and the temporal dependence structure of longitudinal continuous, ordinal, binary, nominal and count data. An MCMC-EM algorithm is developed for parameter estimation.

C0995: **A fuzzy spectral clustering model**
Presenter: **Cinzia Di Nuzzo**, University of Catania, Italy
Co-authors: Giorgia Zaccaria

A new fuzzy approach to the spectral clustering model is introduced. Standard spectral clustering is a technique that exploits the spectral structure of data to partition them into homogeneous groups. Unlike traditional methods such as k-means, spectral clustering does not assume a specific cluster shape and can handle non-linearly separable data. The process involves constructing a similarity matrix, computing the Laplacian matrix, identifying its eigenvectors, and using these eigenvectors to represent the data in a reduced-dimensional space where clustering is more evident. It has been demonstrated that spectral clustering is effective for complex or non-linear structures and can handle high-dimensional data. Integrating this method with a fuzzy approach for clustering has been deemed crucial for maximizing efficiency and coherence in data representation by capturing the intrinsic relationships among data. In the proposed fuzzy method, a least squares approach is used to estimate the model, resulting in a fuzzy Laplacian configuration representative of the entire dataset. The utility of this method is demonstrated through empirical evaluations of synthetic and real-world datasets. The results show the effectiveness of the approach in uncovering complex patterns and providing significant insights. Further developments in a robust framework are also outlined.

CO129 Room S0.13 RECENT ADVANCES IN STATISTICAL MODELING FOR MEDICAL AND SOCIAL DATA Chair: Tiejun Tong

C0330: **An optimal two-step estimation approach for two-phase studies**
Presenter: **Kin Yau Wong**, Hong Kong Polytechnic University, Hong Kong
Co-authors: Qingning Zhou

Two-phase sampling is commonly adopted to reduce costs and improve estimation efficiency. The two-phase study design is considered where the outcome and some cheap covariates are observed for a large cohort in Phase I, and expensive covariates are obtained for a selected subset of the cohort in Phase II. As a result, the analysis of the association between the outcome and covariates faces a missing data problem. The complete case analysis that uses only the Phase II sample is generally inefficient. A two-step estimation approach is developed, which first obtains an estimator using the complete data and then updates it using an asymptotically mean-zero estimator obtained from a working model between the outcome and cheap covariates using the full data. The two-step estimator is asymptotically at least as efficient as the complete-data estimator and is robust to misspecification of the working model. A kernel-based method is proposed to construct a two-step estimator that achieves optimal efficiency, and also develop a simple joint update approach based on multiple working models to approximate the optimal estimator. The proposed method is based on the influence function and is generally applicable as long as the complete-data estimator is asymptotically linear. The advantages of the proposed method are demonstrated over the existing approaches via simulation studies and provide applications to real biomedical studies.

C0296: **Demographic parity-aware individualized treatment rules**
Presenter: **Wen Su**, City University of Hong Kong, Hong Kong
Co-authors: Wenhai Cui, Xiaodong Yan, Xingqiu Zhao

There has been growing interest in developing advanced methodologies aimed at estimating optimal individualized treatment rules (ITRs) in various fields, such as business decision-making, precision medicine, and social welfare distribution. The application of ITRs within a societal context raises substantial concerns regarding potential discrimination. Customized policies, learned from biased data, can inadvertently lead to disparities based on sensitive attributes such as age, gender, or race. To address this concern directly, the concept of demographic parity (DP) is introduced in ITRs. However, estimating an optimal ITR that satisfies the demographic parity definition requires solving a non-convex-constrained optimization problem. To overcome these computational challenges, tailored fairness proxies are employed and inspired by DP to transform them into a convex quadratic programming problem. Additionally, the consistency and convergence rate of the proposed estimator is established. The performance of the proposed method is demonstrated through extensive simulation studies and real data analysis.

C1143: **A hybrid model for zero-inflated proportion data**
Presenter: **Binyan Jiang**, The Hong Kong Polytechnic University, Hong Kong

The aim is to propose a novel hybrid approach to model zero-inflated proportion data by capturing two distinct types of zeros based on a regression framework. The first type of zero is attributed to "missing by chance," arising from random sampling or measurement errors. This type of zero is modelled using a binomial sampling process, accounting for the probability of observing a zero value due to chance. The second type of zero is the so-called "true zero", which is attributed to "unsuitability", reflecting the inherent characteristics of the process or phenomenon under study. This type of zero is handled using a general classification indicator, which separates the observations into suitable and unsuitable groups. The resulting model is, therefore, hybrid, with the regression part utilizing weighted least squares to model the "missing by chance" zeros and the non-zero observations and the classification part to identify the true zeros. A two-stage estimation procedure is employed that involves separating the classification part from the regression part. In particular, the optimal classification rule is identified, and as an illustration, a decision rule based on the nonparametric Nadaraya-Watson estimator is constructed. The consistency of the proposed estimation has been established, and the promising performance of the model has been further demonstrated through simulation and real data analysis.

C0258: **Heterogeneous longitudinal structural equation modeling and variable selection**
Presenter: **Chuoxin Ma**, Beijing Normal University-Hong Kong Baptist University United International College, China

Motivated by the China Health and Retirement Longitudinal study, the primary focus is the association between depressive symptom trajectories and various direct and indirect factors. Since the surveys were taken from about 10,000 households in 150 counties and 450 villages, there may be a substantial heterogeneous grouping effect in the longitudinal measurements of the variables. To address these features, a method is proposed for simultaneous mediator selection, subgroup detection and modeling of the longitudinal trajectories. Longitudinal structural equation modelling (LSEM) with regularization was employed to select the important mediating variables and confounding variables and explore the heterogeneity of the direct effects and mediation effects among different groups. The accuracy of the subgroup detection and variable selection were evaluated in numerical studies. The real analysis suggested that the population can be divided into four groups, and cash assets and medical expenses were important mediators between education level and depressive symptom trajectories.

CO153 Room Safra Lec. Theatre RECENT ADVANCES IN FUNCTIONAL DATA ANALYSIS Chair: Maximilian Ofner

C0220: **An operator-level GARCH model**
Presenter: **Sebastian Kuehnert**, University of Bochum, Germany
Co-authors: Alexander Aue, Gregory Rice, Jeremy Vander Does

Conditional heteroskedastic processes are commonly described by the GARCH model, which has been extensively studied in the uni- and multivariate case and recently also in function spaces. The concept of the functional GARCH model is extended, which has been defined exclusively on function spaces in the point-wise sense. The GARCH model in this article is defined in general, with separable Hilbert spaces, and the GARCH equations consider all the functions. Sufficient conditions for strictly stationary solutions, finite moments and weak dependence are derived, and sufficient and necessary conditions for weak stationarity are discussed. In addition, consistent Yule-Walker estimates with explicit convergence rates are established for the finite-dimensional projections of the GARCH parameters and their entire representation. Finally, the usefulness of the proposed model is demonstrated through a simulation study and a real data example.

C0228: **On the modelling and prediction of high-dimensional functional time series**
Presenter: **Qin Fang**, the University of Sydney, Australia
Co-authors: Jinyuan Chang, Xinghao Qiao, Qiwei Yao

The aim is to propose a two-step procedure to model and predict high-dimensional functional time series, where the number of function-valued time series p is large in relation to the length of time series n. The first step performs an eigenanalysis of a positive definite matrix, which leads to a one-to-one linear transformation for the original high-dimensional functional time series, and the transformed curve series can be segmented into several groups such that any two subseries from any two different groups are uncorrelated both contemporaneously and serially. Consequently, in the second step, those groups are handled separately without the information loss on the overall linear dynamic structure. The second step is devoted to establishing a finite-dimensional dynamical structure for all the transformed functional time series within each group. Furthermore, the finite-dimensional structure is represented by that of a vector time series. Modelling and forecasting for the original high-dimensional functional time series are realized via those for the vector time series in all the groups. The theoretical properties of the proposed methods are investigated, and the finite-sample performance is illustrated through both extensive simulation and two real datasets.

C0664: **Reproducing kernel approach to tomographic data**
Presenter: **Alessia Caponera**, LUISS Guido Carli, Italy

Co-authors: Ho Yun, Victor Panaretos

Many natural phenomena pose challenges wherein the function of interest cannot be directly measured. For instance, the density of a brain cannot be directly measured; rather, it can only be evaluated through 2D sectional images via computerized tomography (CT). Tomography refers to a technique employed to produce sectional images at various orientations using penetrating waves, representing a non-invertible linear operator that maps the original function to a lower-dimensional function, such as positron emission tomography (PET) and quantum state tomography (QST). In such a setup where the true random function is a latent feature, how can their mean function and covariance tensor be estimated using discretized tomographic data? The tomographic operator is considered an operator between reproducing kernel Hilbert spaces (RKHS), and representer theorems are established to address the problem of mean and covariance estimation. The uniform rates of convergence of the estimators are also presented with respect to the observation scheme, evaluating efficiency through simulation results across various tomographic configurations.

C0756: **Heritability modeling of complex functional phenotypes**

Presenter: **Eardi Lila**, University of Washington, United States

Co-authors: Keshav Motwani, Ali Shojaie, Ariel Rokem

Magnetic resonance imaging has played a key role in defining structural and functional measures of brain connectivity. Univariate heritability models have been extensively used to estimate the portion of observed inter-individual differences in connectivity attributable to genetics. However, precisely characterizing how genetics and environmental factors shape these phenotypes remains a challenge due to their complex nature. A novel variance component model designed to enable computationally efficient heritability analysis of complex phenotypes, such as manifold-valued or functional data, is introduced. The proposed model allows for the estimation of primary modes of variation due to genetic and environmental factors, generalizing well-known tools such as tangent and functional principal components analysis. Its application to brain connectivity data reveals that the primary modes of variation, typically characterized by overall increases or decreases in connectivity levels, arise from more complex structural and functional connectomes influenced by genetic and environmental factors.

CO305 Room BH (S) 1.01 Lec. Theatre 1 INFERENCE FOR DEPENDENT DATA Chair: Fabian Mies

C0371: **Series ridge regression for spatial data**

Presenter: **Daisuke Kurisu**, The University of Tokyo, Japan

Co-authors: Yasumasa Matsuda

A general asymptotic theory of series estimators is developed for spatial data collected at irregularly spaced locations within a sampling region. A stochastic sampling design is employed that can flexibly generate irregularly spaced sampling sites, encompassing both pure increasing and mixed increasing domain frameworks. Specifically, the focus is on a spatial trend regression model and a nonparametric regression model with spatially dependent covariates. For these models, L2-penalized series estimation of the trend and regression functions is investigated. Uniform and L2 convergence rates and multivariate central limit theorems are established for general series estimators as the main results. Additionally, it is shown that spline and wavelet series estimators achieve optimal uniform and L2 convergence rates and propose methods for constructing confidence intervals for these estimators. Finally, the dependence structure conditions on the underlying spatial processes include a broad class of random fields, including Levy-driven continuous autoregressive and moving average random fields.

C0534: **On data-driven tuning for truncated realized variations**

Presenter: **B Cooper Boniece**, Drexel University, United States

Co-authors: Jose Figueroa-Lopez, Yuchen Han

Many methods for estimating volatility in the presence of jumps require the specification of tuning parameters for their use in practice. In much of the available theory, tuning parameters are assumed to be deterministic, and their values are specified only up to asymptotic constraints. However, in empirical work and in simulation studies, they are typically chosen to be random and data-dependent, with explicit choices often relying entirely on heuristics. Data-driven fixed-point procedures are discussed for estimating volatility in the presence of jumps. If time permits, some related work on data-driven thresholding procedures in a high-dimensional setting will be discussed.

C0734: **Mixed moving average field guided learning for spatiotemporal data**

Presenter: **Imma Valentina Curato**, TU Chemnitz, Germany

Influenced mixed moving average fields are a versatile modeling class for spatiotemporal data. However, their predictive distribution is not generally known. Under this modeling assumption, a novel spatiotemporal embedding and a theory-guided machine learning approach are defined that employs a generalized Bayesian algorithm to make ensemble forecasts. Lipschitz predictors are employed, and fixed-time and any-time PAC Bayesian bounds are determined in the batch learning setting. Performing causal forecast is a highlight of the methodology as its potential application to data with spatial and temporal short and long-range dependence.

C1320: **Multiscale change detection for non-stationary time series**

Presenter: **Fabian Mies**, Delft University of Technology, Netherlands

Co-authors: Johann Koehne

To increase the power of a CUSUM-based test for structural breaks against short-lived changes, the tests may be localized for various bandwidths and aggregated into a so-called multiscale test statistic. This procedure has recently been shown to possess optimal detection properties against breaks of all sizes and durations. As an intricate multiple-testing procedure, the multiscale test requires control of the finite sample behavior of the process under investigation, which is usually ensured by imposing Gaussianity of the errors or by imposing sub-Gaussian tail bounds. However, the latter can lead to infeasible statistical methods as the sub-Gaussian norm can not be estimated reliably. The method of obtaining a feasible multiscale test is demonstrated via weak convergence arguments. To this end, suitable tightness results are derived for the functional central limit in Holder spaces. Probabilistically, a new kind of restricted weak convergence is discovered, which only holds in the tails of the distribution. The novel theory allows for the optimal detection properties of multiscale tests to be maintained and extended to non-stationary nonlinear time series via a suitable bootstrap scheme.

CO393 Room BH (SE) 1.01 BAYESIAN APPLIED ECONOMETRICS Chair: Kazuhiko Kakamu

C0709: **General Bayesian quantile regression of count via generative modeling**

Presenter: **Yuta Yamauchi**, University of Tokyo, Japan

Co-authors: Genya Kobayashi, Shonosuke Sugasawa

A novel Bayesian framework is presented for estimating quantile regression functions of the discrete response based on the inference of the conditional cumulative distribution of the discrete response. The approach involves the following steps: first, the joint distribution of the discrete response and the covariates is estimated using nonparametric mixture methods. Next, the posterior samples of the conditional quantiles are obtained based on the induced conditional cumulative distribution of the response given the covariates from the estimated joint distribution. Finally, the relationship is represented between the conditional quantile and the covariates using an additive model. The posterior samples of the quantile regression function based on this additive model can be obtained by minimizing the difference between the sampled conditional quantiles and the regression function using a loss function. Through simulation studies, the high flexibility of the method is demonstrated in capturing the relationship between the conditional quantiles and the covariates.

C0790: **Bayesian model averaging for income distributions**
Presenter: **Haruhisa Nishino**, Hiroshima University, Japan
Co-authors: Kazuhiko Kakamu

The aim is to estimate the generalized beta distribution of the second kind (GB2) with four parameters for the income distribution. It is known to be helpful in analyzing income distribution. However, estimating the GB2 by the maximum likelihood estimation has some problems, such as difficulty choosing appropriate initial values, which can lead to unstable estimates. An alternative feasible Bayesian method is to estimate it using the Taylored randomized block Metropolis-Hastings (TaRBMH) algorithm. On the other hand, the GB2 distribution encompasses several three-parameter distributions as special cases, such as the Dagum distribution, the Singh-Maddala distribution, the Beta distribution of the second kind (B2), and the generalized Gamma distribution. Therefore, the Bayesian model averaging is also explored using these three-parameter distributions to estimate income distributions. The simulation study shows that comparing the two Bayesian methods indicates that the latter requires less computation time than the former. The two methods are also applied to actual Japanese equivalent income data (individual and group data) from the comprehensive survey of living conditions. The two methods are thus evaluated, and the characteristics and dynamics of income distributions in Japan are investigated.

C0921: **Estimating spatial decomposition of income inequality via constrained Bayes method**
Presenter: **Yuki Kawakubo**, Chiba University, Japan
Co-authors: Kazuhiko Kakamu

The class of generalized entropy (GE) inequality measures, which includes the widely used Theil index as a special case, has the property of additive decomposability. When the population (entire country) is divided into non-overlapping and exhaustive subpopulations (regions), the GE of the entire country is decomposed into the weighted average of the GE of each region (within-region inequality) and the GE of the mean incomes of the regions (between-region inequality). In this research, the GE of the entire country is estimated, and those of the regions based on grouped data in a way that yields estimates that are compatible with the decomposition. First, the GE of the entire country is estimated by assuming some suitable parametric income distribution. Next, a parametric income distribution is fitted to each region, but as the sample size of each region is often not very large, the parameter vectors are modeled by linking to region-wise auxiliary variables in order to borrow strength from other regions. Based on the model, the GE of each region is estimated using the constrained Bayes method under the constraint that the decomposition holds. The proposed method is applied to the Japanese income data, and the results are compared with those of several existing methods.

C0275: **Bayesian analysis of aging and declining household size on income distribution in Japan**
Presenter: **Kazuhiko Kakamu**, Nagoya City University, Japan

A new dynamic model is proposed to estimate income distributions from grouped data and simultaneously examine the causes of changes in income distribution. Assuming a generalized beta distribution of the second kind as a hypothetical income distribution, a dynamic model is constructed by extending the state-space model. The parameters of the model are estimated using a Markov chain Monte Carlo method. Using Japanese household survey data, the impact of aging and declining household size is examined on income distribution in Japan. The empirical results confirm that both the aging of the population and household size affect income inequality. In particular, both the aging of the population and household size affect an increase in the proportion of lower-income households, affecting income inequality among lower-income households.

CO139 Room BH (SE) 1.02 RECENT ADVANCES IN SYMBOLIC DATA ANALYSIS Chair: Andrej Srakar

C0489: **Covariance estimation for histograms using copulas**
Presenter: **Lynne Billard**, University of Georgia, United States

Histogram-valued data are emerging increasingly often as a consequence of the aggregation of large data sets. One statistic that underpins many methodologies, especially regression and principal component analyses, is the covariance function. Typically, unfortunately, only the marginal distributions are recorded. Therefore, maximum likelihood, inference function for margins, and canonical maximum likelihood estimation methods based on copula functions are proposed. These are then used to obtain estimators of the underlying covariances.

C1009: **Interval-valued models in frequentist and Bayesian schemes**
Presenter: **Abdolnasser Sadeghkhani**, North Carolina Agricultural and Technical State University, United States

Interval-valued data is covered in the multivariate domain, deriving maximum likelihood estimators for parameters and establishing their asymptotic distributions. It introduces a theoretical Bayesian framework for multivariate data, previously applied only to univariate cases. Detailed explanations of the proposed estimators, comparative performance analyses, and validations of their effectiveness through simulations and real-world data analysis are included.

C1006: **Multivariate interval-valued time series data analysis**
Presenter: **S Yaser Samadi**, Southern Illinois University Carbondale, United States

Interval-valued time series (ITS) data are prevalent in many scientific disciplines and applications, including economics, finance, social sciences, and meteorology. Such data arise naturally or through the aggregation of large datasets. Modeling and forecasting of multivariate ITS data have gained significant attention in statistics and related fields. Despite existing efforts in the literature, substantial gaps remain in both theory frameworks and practical applications for appropriately analyzing these data. A new class of interval-valued vector autoregressive (IVAR) models is introduced to capture the cross-dependence dynamics within an ITS vector system. The maximum likelihood estimators of the parameters of the IVAR models are derived, and their asymptotic properties are established. Simulation studies and real data analyses are presented to demonstrate and validate the effectiveness and practical utility of the proposed methodology.

C1197: **Advances in data analysis using aggregated data**
Presenter: **Boris Beranger**, University of New South Wales, Australia
Co-authors: Scott Sisson, Prosha Rahman, Ahmad Hakiim Jamaluddin

The necessity for faster and more efficient statistical modelling techniques has been motivated by the rise of big and complex data. For example, the huge volume of internet data collected on a daily basis implies that simple statistical models cannot be fitted on a regular computer and sometimes even be stored. One strategy is to reduce the amount of data by aggregating it into summaries and to perform an analysis on the summaries themselves. For a general aggregated function, a likelihood-based approach is proposed to fit statistical models defined at the underlying data level. Theoretical guarantees about those maximum likelihood estimators are established, including consistency results for generic continuous aggregation functions. The important yet (almost) unexplored topic of summary design is then dived into. Focusing on the family of (univariate) random bin histogram aggregation functions and developing a methodology to provide some answers to the burning question: how many bins do we need and where to place them? Some simulation experiments are provided to illustrate the insights drawn from the methodology.

CO489 Room BH (SE) 1.06 COMPLEX STATISTICAL METHODS FOR ENERGY POVERTY RELATED ISSUES Chair: Alfonso Carfora

C0691: **How economic backwardness and institutional quality affect energy poverty: Evidence from Italian regional panel data**
Presenter: **Lodovico Santoro**, National Anticorruption Authority (ANAC), Italy

Energy poverty is a major concern in both developing and developed countries, and it is not unusual to find differences within the same country, as in the case of Italy. So far, several studies have used household-level data, but little attention is devoted to analysing the macro-contextual factors in which the usual survey-based determinants of energy poverty operate. The aim is to fill this gap and exploit the great heterogeneity of Italian regions to analyse the effect of economic backwardness and institutional quality on a consensus indicator of energy poverty derived from the EU-SILC survey. The variable calculated as the share of value added from agriculture on total value added is employed as a proxy for economic backwardness, allowing, on the one hand, to make a distinction between the different economic structures of the regions and, on the other hand, to indirectly capture the specific rural dimension of energy poverty. Therefore, by relating the macro and micro dimensions of energy poverty, regressions are carried out on panel data through IV procedures to shed light on the Italian regional macro frameworks in which the phenomenon arises. DISCLAIMER: The opinions expressed here are my own and do not necessarily reflect the view of ANAC.

C0805: **Assessing the targeting effectiveness of fiscal incentives toward energy-poor families**
Presenter: **Leo Fulvio Minervini**, University of Macerata, Italy
Co-authors: Alfonso Carfora

Many EU countries have established long-term tax incentive renovation strategies to support the renovation of their national building stock in line with their national energy and climate plans. Among these countries, Italy offers the most generous incentives. Using the latest data on Italian household expenditure and other socio-demographic characteristics collected by the Household Budget Survey led by ISTAT, the aim is to contribute to the ongoing debate surrounding the efficacy of these incentives in reducing energy poverty. The empirical analysis carried out using a non-parametric counterfactual approach, assesses the impact of fiscal incentives on a subset of four expanded household energy poverty indicators, which are commonly used both in the scientific literature and as official government measures. The evidence, first highlighted by the descriptive analysis and then confirmed by the causal inference perspective through the matching approach, shows a lower prevalence of energy poverty among households that have incurred renovation expenses compared to those that have not. This result indicates that households incurring renovation expenditures, all other conditions being equal, experience a lower average energy poverty condition. Therefore, in this political phase, with significant debate on the effectiveness of current fiscal incentives and pressure to reduce them, our study can help policymakers allocate resources more effectively towards households.

C0845: **Still using solid fuels: Energy poverty in urban areas with different fuel use patterns**
Presenter: **Xinyi Wang**, The Hong Kong University of Science and Technology, Hong Kong
Co-authors: Laurence Delina, Kira Matus, Yueming Qiu

Despite the development of energy transitions and urbanization, traditional solid fuel use persists in urban areas worldwide. However, energy poverty regarding the use of solid fuels in urban areas has been overlooked in energy literature and policy. The household-level demographic and energy data is drawn on in urban China, and a three-way typology of energy poverty is proposed regarding the using solid fuels in urban areas: (1) traditional fuel only, (2) Dirty stacking of Liquified Petroleum Gas (LPG) and traditional fuels, and (3) Dirty stacking of gas infrastructure and traditional fuels. A least absolute shrinkage and selection operator (LASSO) technique is employed to explore the most important drivers for the three fuel use patterns, respectively. It also explores the contribution of the drivers to the models by using Shapley decomposition. Findings suggest that the pattern of "traditional fuel only" fuel use is driven primarily by people's income and education level. The pattern of "LPG stacking with traditional fuels" is driven mainly by living in old-aged housing. The pattern of "gas infrastructure stacking with traditional fuels" is influenced mainly by people's family types. The distinctions between the three fuel use patterns suggest that policies should treat the related populations differently.

C0849: **Explaining SDGs: How predict energy poverty in Italian households**
Presenter: **Giuseppe Scandurra**, Parthenope University of Naples, Italy
Co-authors: Cecilia Camporeale

Energy poverty is a multifaceted and pervasive issue that poses significant challenges to policymakers and affected households. EP, understood as the inability of households to secure sufficient energy services for maintaining an adequate standard of living, makes it difficult to reconcile the objectives of the so-called energy trilemma, the simultaneous pursuit of energy security, environmental sustainability, and socially equitable access to energy sources. As climate change impacts intensify, addressing energy poverty becomes crucial for building resilient communities and fostering a sustainable future. EP measurement is generally based on different class indicators, which can be distinguished from households' perceptions of energy poverty and expenditure-based approaches. A classification algorithm is proposed that identifies Italian families living in energy deprivation using the data from the Household Budget Survey led by the Italian National Institute of Statistics (ISTAT). While exploring various machine learning algorithms, most analysis uses a random forest classifier. The proposed model can accurately classify families in energy poverty, providing valuable insights into the most important variables in predicting the risk of experiencing energy-related distress. In this way, it is possible to identify the levers on which policymakers should act to tackle energy poverty effectively.

C1111: **On a fair definition of energy poverty**
Presenter: **Dusana Dokupilova**, Slovak Academy of Sciences, Slovakia

Energy poverty is a complex problem caused by factors such as low income, high energy expenditures or low energy efficiency of the household (dwelling and appliances). The introduction of measures to reduce the number of households in energy poverty, at the same time as financial resources for such measures are limited, is conditional on a targeted definition that identifies the population groups at risk. There are different approaches to defining the group of people in energy poverty (LIHC, energy expenditure as a percentage of disposable income, etc.) Analyzing individualized data from the housing budget survey combined with field research, it is found that these are not sufficient as there are many households that are not affected by energy poverty and fall into such a group by definition. Or there are households with deep problems, and they are excluded based on the definition. In order to target vulnerable groups as accurately as possible, a new definition is proposed. It is based on the difference between equalized disposable income and energy expenditure, and it is compared to the minimum living wage (based on the national definition).

CO123 Room BH (S) 2.01 ADVANCES IN QUANTITATIVE RISK MANAGEMENT Chair: Hideatsu Tsukahara

C0238: **Forecasting and backtesting gradient allocations of expected shortfall**
Presenter: **Takaaki Koike**, Hitotsubashi University, Japan
Co-authors: Cathy W-S Chen, Edward Meng-Hua Lin

Capital allocation is a procedure for quantifying the contribution of each source of risk to aggregated risk. The gradient allocation rule, also known as the Euler principle, is a prevalent rule of capital allocation under which the allocated capital captures the diversification benefit of the marginal risk as a component of overall risk. This research concentrates on expected shortfall (ES) as a regulatory standard and focuses on the gradient allocations of ES, also called ES contributions. The comprehensive treatment of backtesting the tuple of ES contributions is achieved in the framework of the traditional and comparative backtests based on the concepts of joint identifiability and multi-objective elicitability. For robust forecast evaluation against the choice of scoring function, Murphy diagrams are further presented for ES contributions as graphical tools to check whether one forecast dominates another under a class of scoring functions. Finally, leveraging the recent concept of multi-objective elicitability, a novel semi-parametric model is proposed for forecasting dynamic ES contributions based on a compositional regression model. In an empirical analysis of stock returns, a variety of models are evaluated and compared for forecasting dynamic ES contributions, and the outstanding performance of the proposed model is demonstrated.

C0403: **Dynamic asymmetric tail dependence among multi-asset classes for portfolio management: Dynamic skew-t copula approach**
Presenter: **Toshinao Yoshiba**, Tokyo Metropolitan University, Japan
Co-authors: Kakeru Ito

AC dynamic skew-t copula with cDCC model is proposed to capture dynamic asymmetric tail dependence structure among multi-asset classes (government bonds, corporate bonds, equities, and REITs). The empirical analysis shows that the proposed dynamic AC skew-t copula fits data of multi-asset classes better than other dynamic elliptical copulas, including conventional dynamic skew-t copula, in terms of AIC and BIC. Besides, lower tail dependence coefficients have recently increased compared to upper tail dependence coefficients for all pairs. This indicates that the diversification effect through multi-asset investment has been decreasing, and investors should enhance tail risk management. Furthermore, out-of-sample analysis shows that using dynamic skew-t copula, especially the proposed model, enhances expected shortfall (ES) estimation accuracy and the performance of minimum ES portfolio compared to dynamic t copula and dynamic normal copula. It indicates that capturing dynamic asymmetric tail dependence is crucial for multi-asset investment.

C0472: **Maximum pseudo-likelihood estimation of copula models and moments of order statistics**
Presenter: **Alexandra Dias**, University of York, United Kingdom

It has been shown that despite being consistent and, in some cases, efficient, maximum pseudo-likelihood (MPL) estimation for copula models overestimates the level of dependence, especially for small samples with low levels of dependence. This is especially relevant in finance and insurance applications when data is scarce. It is shown that the canonical MPL method uses the mean of order statistics, and the proposal is to use the median or the mode instead. It is shown that the MPL estimators proposed are consistent and asymptotically normal. In a simulation study, the finite sample performance of the proposed estimators is compared with that of the original MPL and the inversion method estimators based on Kendall's tau and Spearman's rho. In the results, the modified MPL estimators, especially the one based on the mode of the order statistics, have better finite sample performance both in terms of bias and mean square error. An application to general insurance data shows that the level of dependence estimated between different products can vary substantially with the estimation method used.

C0694: **The role of dependences and the moral hazard constraint in optimal reinsurance**
Presenter: **Alexandra Moura**, ISEG Lisbon School of Economics and Management & CEMAPRE - REM, Portugal

The optimal reinsurance problem for several dependent risks is analyzed from the perspective of the direct insurer. The goal is to identify the optimal treaty that maximizes expected utility, assuming independent negotiation of reinsurance for each risk and employing different premium calculation principles for reinsurance. The dependence structure of the risks is general, and the reinsurance premium principles are moment-based. The problem is examined with and without constraints, such as the Lipschitz constraint, which is commonly imposed to prevent moral hazard. Based on the optimality conditions discussed in a prior study, the impact of the Lipschitz constraint on the optimal reinsurance treaty is analyzed. Numerical methods can obtain the optimal treaties for general dependent risks, with or without constraints. Numerical examples are provided, and comparisons of the optimal solutions are conducted.

CO404 Room BH (S) 2.02 EMPIRICAL AND COMPUTATIONAL METHODS FOR FINANCE Chair: Matthias Fengler

C0619: **Artificial neural network small-sample-bias-corrections of the AR(1) parameter close to unit root**
Presenter: **Haozhe Jiang**, Dresden University of Technology, Germany
Co-authors: Ostap Okhrin, Michael Rockinger

An ANN approach is introduced to estimate the autoregressive process AR(1) when the autocorrelation parameter is near one. Traditional OLS estimators suffer from biases in small samples, necessitating various correction methods proposed in the literature. The ANN, trained on simulated data, outperforms these methods due to its nonlinear structure. Unlike competitors requiring simulations for bias corrections based on specific sample sizes, the ANN directly incorporates sample size as input, eliminating the need for repeated simulations. Stability tests involve exploring different ANN architectures and activation functions, as well as robustness to varying distributions of the process innovations. Empirical applications on financial and industrial data highlight significant differences among methods, with ANN estimates suggesting lower persistence compared to other approaches. The same technology may be extended to further research in constructing the estimator of realized higher moments.

C1153: **A non-Gaussian, structure-preserving stochastic volatility and option pricing model in discrete time**
Presenter: **Simon Feistle**, University of Sankt Gallen, Switzerland
Co-authors: Matthias Fengler, Alexander Melnikov

A novel stochastic volatility model is provided based on the autoregressive Gamma process that allows for both a structure-preserving change to the risk-neutral measure and a non-Gaussian distribution for the return innovations. The model employs the Meixner distribution, which enriches the return dynamics with conditional stochastic skewness and kurtosis. A fast and accurate estimation method is proposed by combining the approximate maximum likelihood method of a prior study with a numerical integration technique suitable for highly oscillatory functions. A closed-form discrete-time option pricing formula is derived. The Meixner specification is superior to the benchmark of their family, especially when calibrated to option data.

C1363: **Coherent forecasting of realized volatility**
Presenter: **Marius Puke**, University of Hohenheim, Germany
Co-authors: Karsten Schweikert

The QLIKE loss function is the favored choice in the literature for out-of-sample evaluations of volatility forecasts. However, the state-of-the-art models for forecasting realized volatility (RV), such as the heterogeneous autoregressive (HAR) model, traditionally minimize the squared error loss for in-sample parameter estimation. To the best of knowledge, no prior research has investigated the potential benefits of using the QLIKE loss function also for in-sample parameter estimation within an extensive volatility forecasting exercise. In doing so, the HAR model is embedded within the class of M-estimators, which theoretically justifies the use of the rich class of Bregman loss functions, including, among many others, the squared error and QLIKE loss functions. While asymptotically, the squared error and QLIKE loss functions should yield the same solution, our empirical findings reveal forecast performance gains from a HAR model directly estimated using the QLIKE loss in finite samples. Results are based on the current cross-section of stocks in the Dow Jones index, where the newly proposed HAR model extension demonstrates significant improvements in forecast performance, particularly for short-term horizons.

C1493: **Ex-ante risk timing**
Presenter: **Diego Ronchetti**, Audencia Business School, France

A dynamic ex-ante risk-timing strategy for equity portfolios is introduced, using option-implied risk indicators to determine portfolio weights. This strategy outperforms traditional backward-looking volatility timing methods and beta-pricing models. In the U.S. market over recent decades, it has led to substantial improvements in alphas, Sharpe, Sortino, and Calmar ratios for portfolios sorted by leverage, size, credit rating, and industry, even after accounting for realistic transaction costs and rebalancing frequencies. Findings highlight the importance of options in providing timely insights into firms' capital structures and time-varying return moments, offering a valuable approach to risk management and portfolio allocation.

CO344 Room BH (S) 2.03 CAUSAL INFERENCE IN SUSTAINABLE INVESTING Chair: Serge Darolles

C0661: **Understanding the ESG score effects on stock returns using mediation theory**
Presenter: **Gaelle Le Fol**, Universite Paris - Dauphine, and CREST, France
Co-authors: Serge Darolles, Yuyi He

The influence of ESG scores on stock returns is investigated. The objective is to use these scores to study the channels through which the ESG effects are transmitted (if any). The theory of the ESG transmission mechanism comes through essentially two channels: the "investor demand channel" and the "fundamentals or profitability channel." However, in practice, these effects are difficult to identify and quantify. To tackle this problem, a causal mediation analysis is proposed. This approach not only allows for seeing whether common ESG scores can predict future returns but also to identify the channels at play. The results show that ESG scores have a real effect on stock returns and that the transmission channels are not the same depending on the awareness of institutional investors (ESG-aware vs ESG-unaware) as well as on the dimension - either E, S, G or global ESG - of focus. Results show that ESG scores indeed contain information that can be exploited by asset managers.

C0735: **An anatomy of decarbonizing firms**
Presenter: **Guillaume Coqueret**, Emlyon Business School, France
Co-authors: Thomas Giroux, Borui Qiu

The purpose is to shed light on the drivers of decarbonization in the cross-section of global publicly-listed firms. Both panel models and feature importance from random forests show that the single best predictors of emission and intensity reduction are proxies of size, whether they are past capitalization, revenue, or emissions. It is found that many other factors influence this, including macroeconomic (country-specific) variables and prior financial decisions. The results indicate that the linear extrapolation method can be improved for the estimation of corporate greenhouse gas trajectories. Asset managers seeking net-zero targets can benefit from the higher accuracy of the approach. A causal analysis of the exogenous macroeconomic events that nudge companies to accelerate their decarbonization path is further conducted.

C0888: **Climate regulatory risk and technological opportunities**
Presenter: **Helene Mathurin**, ESSEC Business School, France

A consumption-based asset pricing model is introduced with long-run risk in which environmental policies stimulate research and development efforts, thereby increasing the probability of breakthrough green innovation. This model features two separate but correlated premia for transition risks: a premium for regulation risk and a premium for technological innovation. The model is calibrated, and the complex interaction between these two premia is illustrated. Finally, the model's predictions are tested by building new measures of these premia, which use press releases to proxy firms' exposure to innovation and to regulatory risk.

C0663: **Betting against sustainability: Evidence from US equity short selling activity**
Presenter: **John Coadou**, Amundi Asset Management / Université Paris Dauphine PSL, France
Co-authors: Serge Darolles

Institutional investors have increasingly incorporated environmental, social, and governance (ESG) factors into their investment policies. The integration of sustainability considerations into their decision-making processes has led to heightened demand for non-sin stocks, influencing asset prices. The literature suggests that the notable returns observed from sustainability investing are primarily driven by the significant price impact of ESG-related capital flows. Given that these investors operate under various constraints - such as the "long-only" mandate and ESG portfolio target scores - alternative investment firms may seek to benefit from a subsequent ESG premium. Monthly short sale data from the IHS Markit database is used to investigate whether hedge funds "bet against ESG" to benefit from this premium and generate positive arbitrage returns in flowing such strategies.

CO384 Room BH (S) 2.05 LATENT VARIABLE MODELLING Chair: Roberto Casarin

C0202: **Media bias and polarization through the lens of a Markov switching latent space network model**
Presenter: **Antonio Peruzzi**, Ca' Foscari University of Venice, Italy
Co-authors: Roberto Casarin, Mark Steel

News outlets are now more than ever incentivized to provide their audience with slanted news, while the intrinsic homophilic nature of online social media may exacerbate polarized opinions. A new dynamic latent space model is proposed for time-varying online audience-duplication networks, which exploits social media content to conduct inference on media bias and polarization of news outlets. The model contributes to the literature in several directions: 1) The model provides a novel measure of media bias that combines information from both network data and text-based indicators; 2) The model is endowed with Markov-switching dynamics to capture polarization regimes while maintaining a parsimonious specification; 3) The contribution to the literature is on the statistical properties of latent space network models. The proposed model is applied to a set of data on the online activity of national and local news outlets from four European countries in the years 2015 and 2016. Evidence of a strong positive correlation is found between the media slant measure and a well-grounded external source of media bias. In addition, insight is provided into the polarization regimes across the four countries considered.

C0725: **Spatial heterogeneity in the effects of covariates on the distribution of income using Bayesian nonparametric methods**
Presenter: **Ziyou Wang**, Kings College London, United Kingdom
Co-authors: Maria Kalli

The purpose is to extend a prior study that developed normalized latent measure factor models. It models a collection of distributions using linear combinations of latent factor measures. These latent factor measures are modelled as infinite mixtures, the weights of which are interpretable as characteristic traits shared by different distributions. The prior study focuses on estimating the density of personal income for the public use microdata sample (PUMS) classifications of the state of California. The mentioned study is extended by looking at how personal income is distributed in the other US states and by considering multiple discrete value covariates, such as gender and race. This allows gaining valuable insights into how the covariates affect the income distribution and how these effects vary over space.

C0853: **A data-rich yield curve factor model**
Presenter: **Andrea Trovato**, Ca Foscari University, Italy
Co-authors: Roberto Casarin, Davide Raggi

A data-enriched term-structure model for interest rates is presented following a general term-structure architecture based on non-arbitrage arguments. The state-space model includes the traditional Vasicek measurement equation with the spot rate as a state process, which is augmented with further measurement equations on bonds at different maturities. The model incorporates additional exogenous variables related to monetary policy, equity market volatility and macroeconomic fundamentals. All the measurement equations are driven by the state process, and the augmentation allows for an improved estimation of the latent interest rate. A Bayesian inference framework is proposed based on an efficient posterior approximation procedure. An innovative and more interpretable way is presented to estimate the risk premium as a combination of explicit sources of risk. The model is applied to the US yield curve sampled at a monthly frequency from December 2000 up to the end of May 2024. The augmented framework and the Bayesian inference allow for generating projections of the yield curve under different market scenarios. The simulation provides great support to decisions of tactical asset allocation since it allows for evaluating the coherence of the portfolio allocation with the predominant regime implied in the market yield curve.

C1601: **Sparse dynamic Bayesian graphical models**
Presenter: **Gregor Kastner**, University of Klagenfurt, Austria
Co-authors: Luis Gruber, Matteo Iacopini

Gaussian graphical models have become a staple in statistical modelling and for estimating partial correlation networks. The baseline approach is extended to a time series framework by introducing temporal dependence for the entries of the precision matrix. In order to conduct statistical inference, a fully Bayesian approach is adopted, relying on global-local shrinkage priors to deal with high-dimensional data and mitigate (temporal) overfitting. The framework has several special cases of interest, including a variant of the standard Bayesian graphical lasso. Closed-form recursions for the filtering and smoothing distributions are obtained. These results are exploited to design a simple yet efficient blocked Gibbs sampler for posterior inference. An interweaving strategy is applied to enhance the mixing of the sampler. As a by-product, the proposed method allows for estimating a time series of (sequentially dependent) networks from partial correlations among the variables in the system. Using synthetic and real data, the performance of the model is investigated in comparison to standard Bayesian and frequentist benchmarks in terms of both covariance matrix estimation and graphical structure learning.

CO248 **Room BH (SE) 2.01** **RECENT ADVANCES IN SAMPLE SELECTION MODELS** **Chair: Francisco Javier Rubio**

C0612: **A generalized Heckman model with varying sample selection bias and dispersion parameters**
Presenter: **Wagner Barreto-Souza**, University College Dublin, Ireland
Co-authors: Fernando Souza, Marc Genton

A generalization of the Heckman sample selection model is proposed by allowing the sample selection bias and dispersion parameters to depend on covariates. It is shown that the non-robustness of the Heckman model may be due to the assumption of the constant sample selection bias parameter rather than the normality assumption. The proposed methodology allows understanding which covariates are important to explain the sample selection bias phenomenon rather than only forming conclusions about its presence. Further, the approach may attenuate the non-identifiability and multicollinearity problems faced by the existing sample selection models. The inferential aspects of the maximum likelihood estimators (MLEs) for the proposed generalized Heckman model are explored. More specifically, it is shown that this model satisfies some regularity conditions such that it ensures consistency and asymptotic normality of the MLEs. Proper score residuals for sample selection models are provided, and model adequacy is addressed. Simulated results are presented to check the finite-sample behavior of the estimators and to verify the consequences of not considering varying sample selection bias and dispersion parameters. It is shown that the normal assumption for analyzing medical expenditure data is suitable and that the conclusions drawn using the approach are coherent with findings from prior literature.

C0641: **Bayesian variable selection in sample selection models using spike-and-slab priors**
Presenter: **Adam Iqbal**, Durham University, United Kingdom
Co-authors: Emmanuel Ogundimu, Francisco Javier Rubio

Sample selection models represent a common methodology for correcting bias induced by data missing not at random. It is well known that these models are not empirically identifiable without exclusion restrictions. In other words, some variables predictive of missingness do not affect the outcome model of interest. The drive to establish this requirement often leads to the inclusion of irrelevant variables in the model. A recent proposal uses adaptive LASSO to circumvent this problem, but its performance depends on the so-called covariance assumption, which can be violated in small to moderate samples. Additionally, there are no tools yet for post-selection inference for this model. To address these challenges, we propose two families of spike-and-slab priors to conduct Bayesian variable selection in sample selection models. These prior structures allow for constructing a Gibbs sampler with closed-form conditionals, which is scalable to the dimensions of practical interest. We illustrate the performance of the proposed methodology through a simulation study and present a comparison against adaptive LASSO and stepwise selection.

C0868: **Type II Tobit sample selection models with Bayesian additive regression trees**
Presenter: **Eoghan O Neill**, Erasmus University Rotterdam, Netherlands

The purpose is to introduce type II Tobit Bayesian additive regression trees (TOBART-2). BART can produce accurate individual-specific treatment effect estimates. However, in practice, estimates are often biased by sample selection. The type II Tobit sample selection model is extended to account for nonlinearities and model uncertainty by including sums of trees in both the selection and outcome equations. A Dirichlet process mixture distribution for the error term allows for departure from the assumption of bivariate normally distributed errors. Soft trees and a Dirichlet prior to splitting probabilities improve the modelling of smooth and sparse data-generating processes. A simulation study and an application to the RAND Health Insurance Experiment data set are included.

C0274: **Model selection under sample selection and model misspecification**
Presenter: **Francisco Javier Rubio**, University College London, United Kingdom
Co-authors: Adam Iqbal, Emmanuel Ogundimu

The effects of model misspecification and sample selection in Bayesian variable selection are discussed, focusing on the formulation of local and non-local priors for both the outcome and selection processes. Strategies for calibrating these priors will also be discussed. Additionally, we address computational challenges related to the non-concavity of the log-likelihood function. Theoretical and empirical results will be presented to illustrate the impact of model misspecification, along with a brief real-data example.

CO144 **Room BH (SE) 2.05** **EMPIRICAL ANALYSIS OF CLIMATE CHANGE** **Chair: Marina Friedrich**

C0401: **Global climate anomalies and economic welfare: Evidence from panel of US counties**
Presenter: **Lotanna Emediegwu**, Manchester Metropolitan University, United Kingdom

An extensive dataset is meticulously compiled, encompassing a range of economic indicators for all counties within the continental United States. Employing a sophisticated econometric approach, the causal impact of the El Nino Southern Oscillation (ENSO) phenomenon on the U.S. economy is investigated. The findings reveal a notable and statistically significant negative correlation between ENSO and the rate of economic growth, both in aggregate and across specific sectors of the economy. Furthermore, distinct effects of the two ENSO regimes are identified, highlighting that El Nino exerts a more pronounced detrimental effect on the U.S. economy compared to La Nina. These results exhibit geographical variations and remain robust across multiple sensitivity checks, minimizing the likelihood of significant deviations from the reported findings.

C0587: **Droughts, migration and population in Kenya**
Presenter: **Melanie Gittard**, Stanford University, United States

Since 2000, Kenya has faced increasing drought frequency, significantly impacting agriculture and driving labor migration. Strategic migration patterns are investigated among farmers and pastoralists in response to repetitive droughts. Using fine-grained data that captures short-distance migration and heterogeneity, it combines satellite-based daily rainfall data with exhaustive censuses from 1989, 1999, and 2009. A two-way fixed-effect model exploits spatial variation in drought frequency across 2,518 sub-locations, comparing demographic growth by drought frequency over each decade. First, increased drought frequency triggers out-migration, as one additional drought decreases demographic growth by 1.7 p.p, equating to a 1% population decline. This effect is consistent within the [15; 65] age group, confirming migration as the driving factor. The main contribution is identifying different migration strategies across livelihoods. Rural areas dominated by pastoral activities experience significant out-migration, leading to a rural-rural shift from pastoral to agriculture-oriented regions. Herders' migration shows little heterogeneity, suggesting entire households migrate, consistent with migration as a last resort. Agricultural rural areas display significant heterogeneity, with educated individuals migrating while uneducated individuals remain in affected areas. The importance of detailed data is underscored for understanding diverse migration strategies.

C0593: **Common persistent cycles**
Presenter: **Anthoulla Phella**, University of Glasgow, United Kingdom
Co-authors: Vasco Gabriel, Luis Filipe Martins

Paleoclimatic time series such as ice volume, CO2 emissions, and temperatures have very similar time series patterns during ice ages and interglacial periods, particularly common persistent long cycles. Their study is revisited using a dataset that ranges from the period -145,000 to the present (in 1000-year intervals), contributing to the empirical debate on how temperatures are determined by CO2 and orbital variables. A novel state-space approach is developed that offers a simple, reliable and robust characterization of common persistent cycles, which are present in two or more time series. The first contribution is methodological. To explore the properties of the model that capture the dynamics of a time series with persistent cycles fully, the single-observed state space representation is started. The variable modelled is decomposed into two components, each depending on one or more state variables: noise and cycle. The first moments of the model's variable, the associated Kalman filter, and the properties of the (Q)MLE estimator are derived. As an extension, the properties of a multivariate version of the previous model are studied where more than one variable common persistent cycles are shared and are affected distinctively by the short-term dynamics. The model is then empirically found to fit well the observed paleoclimatic data, providing a simple alternative to the methods developed in a recent study.

C0965: **Forecasting the atmospheric ethane burden above the Jungfraujoch with Bayesian and frequentist methods**
Presenter: **Marina Friedrich**, VU Amsterdam, Netherlands
Co-authors: Yuliya Shapovalova, Karim Moussa, David van der Straten

Short-lived climate forcers are broadly divided into methane and non-methane volatile organic compounds (NMVOC). They affect the climate and are often also air pollutants. Ethane is the most abundant NMVOC in the atmosphere, sharing important emission sources with methane. The main sources of ethane are anthropogenic, while methane has natural and anthropogenic sources. Understanding trends in atmospheric ethane is crucial to better constrain the anthropogenic sources of methane, in particular from the oil and gas industry. While previous studies focus on analyzing past trends, the aim is to forecast the atmospheric ethane burden above the Jungfraujoch (Switzerland), using both Bayesian and frequentist methods. Since ethane measurements can only be taken under clear-sky conditions, a substantial fraction of the data (around 75%) is missing. The presence of missing data complicates the analysis and limits the availability of appropriate forecasting methods. Three distinct approaches for forecasting ethane time series are employed: 1) state-space modeling; 2) kernel regression which has previously been used for trend analysis in ethane time series; 3) a Gaussian process regression, which can be seen as Bayesian non-parametric regression, with interpretable compositional kernel and a spectral mixture kernel. These three approaches excel in different aspects of time series forecasting, such as flexibility, interpretability, uncertainty quantification, and handling missing data.

CO039 Room BH (SE) 2.09 TWEETS, INFLATION AND MACROECONOMIC POLICIES Chair: Etsuro Shioji

C0676: **How did people tweet against inflation in Japan**
Presenter: **Toshitaka Sekine**, Hitotsubashi University, Japan

During the chronic deflation era starting in the 1990s, Japanese inflation expectations were said to be firmly anchored at a very low level, say, around zero. These expectations seemed to have become something like the social norm. Households were quite against any price hikes, and as a consequence, firms hesitated to raise their prices; when they raised prices, they apologized for their misbehavior. People not only expected that prices would not increase but also believed that prices should not increase. A natural language processing technique is applied to tweets that comment on price changes to uncover whether there has been any change in households' sentiments against price hikes in recent years.

C0787: **New approach to estimating the productivity of public capital: Evidence from 22 OECD countries**
Presenter: **Hiroshi Morita**, Tokyo Institute of Technology, Japan

Investigating the productivity of public capital is a long-standing issue in one strand of macroeconomic literature. A new approach is developed to estimate the output elasticity of public capital in a vector autoregressive model with identification restrictions derived from the theoretical model. The empirical analysis of 22 OECD countries for the period 1960- 2019 reveals that public capital accumulation has a positive effect on GDP both in the short-run and long-run horizons in all countries, supporting both demand-stimulating and growth-enhancing effects. Furthermore, the estimated output elasticity of public capital lies within a reasonable range between 0 and 0.5 and, as in the literature, shows substantial differences across countries. Therefore, it can be concluded that the proposed methodology is valid for studying public capital productivity.

C1185: **State-dependency of fiscal price puzzle**
Presenter: **Naoto Soma**, Yokohama National University, Japan

The aim is to examine the state-dependent nature of the fiscal price puzzle, where expansionary fiscal policy doesn't increase prices as traditional theory predicts. Extending a prior study, a state-dependent local projection model is used to analyze price responses to fiscal shocks during economic expansions and recessions. The findings show significant state dependence: during expansions, government spending increases do not raise prices, and total factor productivity (TFP) rises, which is consistent with the puzzle. However, during recessions, spending increases cause inflation without TFP growth. This suggests that the channel proposed by the aforementioned study may only apply during expansions, while recessionary effects align with conventional theory.

C0212: **Public investment news shocks: A text-based index**
Presenter: **Etsuro Shioji**, Chuo University, Japan

News regarding future public investment enters the private sector information set well in advance of policy implementation. Failing to consider this could lead to an inaccurate measurement of the policy's impact. The purpose is to develop a method to construct a public investment news indicator through text analysis of newspaper articles. The method is applied to the case of Japan. The resulting index is used as an external instrument to estimate the impact of a public investment news shock on the aggregate economy.

CO359 Room BH (SE) 2.12 TOPICS IN PANEL DATA MODELS AND THEIR APPLICATIONS Chair: Chaowen Zheng

C0466: **Dynamic quantile panel data models with interactive effects**
Presenter: **Chaowen Zheng**, University of Southampton, United Kingdom

A simple two-step procedure is proposed for estimating the dynamic quantile panel data model with unobserved interactive effects. Factors are first consistently estimated via an iterative principal component analysis to account for the endogeneity induced by the correlation between factors and lagged dependent variable/regressors. In the second step, a quantile regression is run for the augmented model with estimated factors and estimated slope parameters. In particular, a smoothed quantile regression analysis is adopted where the quantile loss function is smoothed to have well-defined derivatives. The proposed two-step estimator is consistent and asymptotically normally distributed but subject to asymptotic bias due to the incidental parameters. The split-panel jackknife approach is then applied to correct the bias. Monte Carlo simulations confirm that the proposed estimator has good finite sample performance. Finally, the usefulness of the proposed approach is demonstrated with an application to the analysis of bilateral trade for 380 country pairs over 59 years.

C0711: **Estimation of dynamic panel models with interactive effects**
Presenter: **Wenting Wang**, University of York, United Kingdom
Co-authors: Jia Chen, Yongcheol Shin

An internal instrument variable estimation method is presented for dynamic panel data models with unobserved common factors. The main idea involves using a cross-sectional average of regressors to project out common factors potentially emerging in both regressors and dependent variables. The defactorized regressors and their lags are subsequently utilized as instruments, followed by the implementation of the two-stage least squares (2SLS) on the defactorized outcome equation. Theoretical findings include the derivation of consistency and asymptotic normality of individual and mean group estimators in heterogeneous slope model, as well as the pool estimator in homogeneous slope model. These findings are supported by Monte Carlo simulations, which demonstrate the satisfactory performance of the proposed estimators in finite samples.

C1269: **Network analysis of business cycle synchronization using simultaneous equations with interactive effects**
Presenter: **Ting Xie**, University of York, United Kingdom

The simultaneous equation panel data model is developed to jointly accommodate the local spatial correlation and the global factor dependence as well as parameter heterogeneity. The regional network analysis is then proposed to examine the diffusion impacts of the trade and financial intensities on business cycle synchronization (BCS). The CCEX-2SLS approach is applied to the dataset consisting of 136 country pairs in the 17 OECD countries over 1995-2019 and convincingly unveils that both trade and financial intensities can boost the BCS and the effects diffuse from high-income regions operating on or near the frontier to low-income regions. This suggests that policies to reduce trade barriers and encourage bilateral liberalization appear better suited to improve the BCS of net shock receivers, whilst policies to attract more investments are appropriate to their transmitters. In this regard, the importance of investing funds is stressed in peripheral regions to amplify regional BCS.

C1293: **Spatial-temporal synthetic error model of causal analysis with application to policy causal effect evaluation**
Presenter: **Yan Zhang**, University of Southampton, United Kingdom

Causal analysis of spatial-temporal data is challenging owing to spatial-temporal interactions. The synthetic control method (SCM) is popular in estimating the causal effect of a given intervention on a single or a small number of units in a non-spatial panel data setting by weighted averaging of the control units to balance the outcomes and covariates of the treated unit. Inspired by the ideas of synthetic control method and spatial-temporal models, a spatial-temporal synthetic error model (STSEM) is proposed as a new framework of linear spatial-temporal causal inference model to infer the causal effect of some given intervention on the metric that is of interest for spatial-temporal data, with its synthetic weights determined by LASSO regression. Asymptotic properties of the proposed model are established, followed by which the significance of the causal effect can be tested. In addition, its performance is also compared in causal effect inference with the traditional SCM, the augmented SCM (ASCM), and a simplified STAR-PLR model in a simulation study and an empirical study, in which the causal effect of the Kansas tax cut on its GDP is demonstrated for inference.

CC491 Room S-1.01 TEXT DATA AND RELATED TOPICS Chair: Etienne Marceau

C1612: **Topic Model for multiple supervised information based on non-linear functions**
Presenter: **Kotono Waki**, Doshisha university, Japan
Co-authors: Shintaro Yuki, Hiroshi Yadohisa

In marketing, purchase history comprises the number of purchases made by each consumer for each product and is characterised by high dimensionality and extreme sparseness. In many cases, supervisory information, such as consumer attribute data and service evaluations, corresponds to purchase history. Supervised LDA (SLDA) is a method for learning consumer requirements from purchase history and interpreting them based on supervisory information. Supervisory information is crucial for understanding consumer requirements, and SLDA can accommodate a single piece of supervisory information. However, in cases involving multiple pieces of supervisory information, SLDA cannot capture their correlated structure. Additionally, SLDA assumes a linear relationship between consumers and topics during learning, which limits the expressive ability of the model. By contrast, an embedded topic model can estimate topics using nonlinear functions. The aim is to develop a nonlinear topic-learning method that considers the correlated structure of multiple pieces of supervisory information based on the abovementioned methods. This approach enables a more precise understanding of consumer requirements.

C1299: **Time-varying weighted latent Dirichlet allocation**
Presenter: **Louisa Kontoghiorghes**, Kings College London, United Kingdom
Co-authors: George Kapetanios

A time-varying weighted latent Dirichlet allocation (wLDA), a generative probabilistic topic modeling method that can track the evolution of topics in a series of documents, is introduced. This approach combines the latent Dirichlet allocation (LDA) with time-varying weights, estimating the model's corpus-dependent parameters with the weighted log-likelihood. In the time-varying wLDA, the estimation of time-varying topics is done using the most recent documents according to the time-varying weights in each time index, the so-called rolling window estimation. This approach accounts for the importance of documents, giving terms in more influential documents greater contribution to the topic distribution estimation. In addition, this methodology addresses topic estimation in the presence of an imbalanced number of documents in each set within the corpus. The methodology is applied to assess the topic evolution of the abstracts of a scientific conference.

C1390: **Application of latent semantic scaling to high-dimensional text data for personality assessment**
Presenter: **Yoshito Tan**, University of Tokyo, Japan
Co-authors: Keishi Nomura, Kensuke Okada

Interest in the forced-choice assessment format in high-stakes contexts, such as personality assessment in personnel selection, has been increasing because it can mitigate social desirability bias by matching the social desirability levels of the personality trait words being compared. However, obtaining social desirability ratings beforehand is time- and cost-intensive. To address this issue, leveraging the strong correlation between emotional valence and social desirability of trait words and using latent semantic scaling (LSS) to scale unidimensional valences as a proxy is proposed. As a semi-supervised sentiment analysis technique, LSS is interpretable and cost-efficient because it combines positive and negative seed words for weak supervision and unsupervised learning of word embeddings from high-dimensional text data. Given pre-trained embeddings, LSS can be seen as a simple linear model that applies the cosine similarity matrix of embeddings as a linear operator and maps the initial valences of seed words onto the scaled valences of other words. To evaluate the usefulness of the LSS-based valences, the extent to which they predicted the social desirability scores of trait words in the Big-Five personality assessment is examined. Medium-to-strong correlations are observed between them. These results imply that the proposed method contributes considerably to efficient forced-choice format construction.

C0670: **The power of visuals: Using social media images for financial sentiment analysis**
Presenter: **Erik-Jan Senn**, University of St. Gallen, Switzerland
Co-authors: Francesco Audrino

Financial sentiment analysis focuses mainly on text data. However, the importance of visual information from images has increased over the last decades, especially on social media. The objective is to investigate whether visual information influences the sentiment of retail investors and improves financial forecasting. The proposed sentiment model is based on visual information for stock-related posts on the social media platform StockTwits. The images are processed by a computer vision model and classified using user-labelled sentiment. The empirical analysis shows how visual sentiment impacts the classification performance of standard text-based models. In a financial forecasting application, the value of visual information is evaluated for financial variables such as realized volatility.

CC450 Room S-1.04 CONTRIBUTIONS IN CAUSAL INFERENCE Chair: Massimo Cannas

C1268: **Causal discovery for linear acyclic models with gaussian noise using ancestral relationships**
Presenter: **Ming Cai**, Graduate School of Informatics, Kyoto University, China
Co-authors: Penggang Gao, Hao Wang, Hisayuki Hara

The PC algorithm proposed relies solely on the faithfulness assumption of the causal model and identifies causal structures up to the Markov equivalence class. In contrast, LiNGAM added the assumption of the linearity and continuous non-Gaussianity of disturbances to the causal model, which allows for the complete identification of the causal DAG. By integrating the strengths of both the PC algorithm and LiNGAM, PC-LiNGAM makes it possible to identify the causal structure up to the distributed equivalence pattern beyond the Markov equivalence class, even in the presence of some Gaussian disturbances. However, PC-LiNGAM suffers from high computational complexity, which is factorial about the number of variables in the worst case. A new algorithm that significantly reduces the time complexity for learning distributed equivalence patterns is introduced. The main idea of the proposed method is to use the causal ancestor finding of Maeda and Shimizu, which is generalized to include Gaussian disturbances.

C1317: **Handling endogenous regressors in quantile regression models: Copula approach without instruments**
Presenter: **Rouven Haschka**, Zeppelin University Friedrichshafen, Germany

Endogeneity in quantile regression models has not yet received much attention in the literature. In order to handle regressor endogeneity, only instrument-based approaches are used. Seeing that instruments are often weak or unavailable, this article proposes a generalisation for the instrument-free Gaussian copula-based endogeneity correction to quantile regression models. For this purpose, two estimators are developed. First, a full maximum likelihood estimator based on directly maximising the likelihood derived from the joint distribution of explanatory variables and errors given the assumption that errors follow an asymmetric Laplace distribution. Second, a Bayesian estimator that is based on a decomposition of errors into an unobserved exponential variable and a structural normal part. By assuming that endogeneity comes from the normally distributed part, the copula endogeneity correction by control functions is used so that the model can be estimated by efficient Gibbs sampling. Identification assumptions are derived and shown under which conditions these are fulfilled. Moreover, Monte Carlo simulation results are provided to examine and compare the finite sample performances of the two abovementioned estimators and demonstrate their superiority to instrumental variable estimation in quantile regression models.

C1529: **Robust synthetic control method for data with outliers**
Presenter: **Riku Yamashita**, Doshisha University, Japan
Co-authors: Kei Tsubotani, Kensuke Tanioka, Hiroshi Yadohisa

In practical applications, accurately estimating the effects of interventions on outcomes is important. The synthetic control method is widely applied in various fields, such as economics and political science, particularly when randomized experiments are not feasible. This method produces a potential outcome that would have been observed for the treated unit in the absence of treatment by creating a weighted combination of control units that closely matches the characteristics of the treated unit before the intervention. However, the standard synthetic control approach is sensitive to outliers in the treated unit, leading to biased estimates. To address this problem, an outlier-robust synthetic control method that replaces the L2 norm is proposed in the objective function with the L1 norm. This approach can reduce the influence of outliers and improve the robustness of the proposed method. This modification enhances the accuracy of the treatment effect, and the applicability of the synthetic control method may extend to a wider range of real-world scenarios. The effectiveness of the proposed method is demonstrated through simulations and empirical examples, highlighting its practical utility in contexts where data anomalies or outliers are present.

C0607: **Analyzing the impact of social media on the trading behavior of retail investors**
Presenter: **Jule Schuettler**, University of St.Gallen, Switzerland
Co-authors: Enrico De Giorgi, Christoph Hirt

The purpose is to investigate the causal relationship between social media attention and retail investors trading behavior. Using data from WallstreetBets to measure social media attention and from the online brokerage Robinhood to track trading activities, a potential outcomes framework is applied to evaluate the impact of social media on investors' trading behavior. The causal estimation strategy allows for the analysis of the extent to which social media influences retail trading patterns.

CC485 Room K0.16 NETWORKS AND GRAPHICAL MODELS Chair: Andrew Wood

C1578: **Portfolio selection with complex network analysis**
Presenter: **Roope Rihtamo**, University of Turku, Finland
Co-authors: Joni Virta, Harto Saarinen

Complex network analysis methods are used to derive optimal asset weights in a standard portfolio optimization problem. In the proposed empirical approach, different assets are represented by nodes that are linked together to form a network of assets. The links are based on similarity measures that are designed to capture the covariance structure of returns in the opportunity set. Different aspects of return similarity are addressed with measures motivated by financial theory as well as measures motivated by complex network analysis. New and enhanced measures of asset centrality are developed - a key aspect of portfolio diversification. The selected similarity measures and the selected investment opportunity set affect the structure of the formed network which, in turn, defines the optimal weights. Hence, the proposed methods offer novel extensions to the use of complex network analysis methods in financial economics. The proposed methods have straightforward interpretations rooted in theoretical asset pricing and can be easily implemented in various markets.

C1438: **A network approach to macroprudential buffers**
Presenter: **Yuliang Zhang**, LSE, United Kingdom

Network modelling of systemic risk is used to set macroprudential buffers from an operational perspective. The focus is on the countercyclical capital buffer, an instrument designed to protect the banking sector from periods of excessive growth associated with a build-up of system-wide risk. An indicator of financial vulnerability is constructed with a model of fire sales, which captures the spillover losses in the system caused by deleveraging and joint liquidation of illiquid assets. Using data on the U.S. bank holding companies, it is shown that the indicator is informative about the build-up of vulnerability and can be useful for setting the countercyclical capital buffer.

C1535: **Quality, location, and coffee price returns: A high-dimensional CoVaR-copula network analysis**
Presenter: **Luis Fernando Melo Velandia**, Banco de la Republica and Los Andes University, Colombia
Co-authors: Mahicol Stiben Ramirez-Gonzalez

Daily coffee price returns are explored over a two-decade period across different varieties in the U.S., France, and Germany markets. The analysis examines how the quality of coffee beans and their trading locations impact network connections using a high-dimensional CoVaR-Copula network.

The findings reveal that quality significantly influences market connectivity, particularly with premium beans strengthening links between the U.S. and German markets. Colombian coffees consistently cluster as high-quality. Conversely, in France, spatial trading dynamics take precedence over quality considerations, with low-quality beans primarily affecting domestic market risk.

C1499: **A review in Bayesian structure learning in Gaussian graphical models**
Presenter: **Lucas Vogels**, University of Amsterdam, Netherlands
Co-authors: Reza Mohammadi, Marit Schoonhoven, Ilker Birbil

Gaussian graphical models provide a powerful framework to reveal the conditional dependency structure between multivariate variables. The process of uncovering the conditional dependency network is known as structure learning. Bayesian methods can measure the uncertainty of conditional relationships and include prior information. However, frequentist methods are often preferred due to the computational burden of the Bayesian approach. Over the last decade, Bayesian methods have seen substantial improvements, with some now capable of generating accurate estimates of graphs up to a thousand variables in mere minutes. Despite these advancements, a comprehensive review or empirical comparison of all recent methods has not been conducted. The aim is to delve into a wide spectrum of Bayesian approaches used for structure learning and evaluate their efficacy through a comprehensive simulation study. The application of Bayesian structure learning is also demonstrated in a real-world data set, and directions for future research are provided. An exhaustive overview of this dynamic field is given for newcomers, practitioners, and experts.

CC494 Room S0.03 ANALYSIS OF CATEGORICAL DATA Chair: Frederic Ferraty

C1127: **Hierarchical imputation of categorical variables in the presence of systematically and sporadically missing data**
Presenter: **Shahab Jolani**, Maastricht University, Netherlands

In the development of prediction models, data are often combined from different sources, known as individual participants data (IPD) sets. A specific challenge in analysing IPD sets is the presence of systematically missing data when certain variables are not measured in some studies and sporadically missing data when measurements of certain variables are incomplete across different studies. Multiple imputation (MI) is among the better approaches to deal with missing data. However, MI of clustered data, such as IPD meta-analysis, requires advanced imputation routines that preserve the hierarchical structure of data and accommodate both systematically and sporadically missing data. A new class of hierarchical imputation methods have been recently developed within the MICE framework tailored for continuous variables. The extension of these methods to categorical variables is discussed, accommodating the simultaneous presence of systematically and sporadically missing data in nested designs with arbitrary missing data patterns. To address the challenge of the categorical nature of the data, an accept-reject algorithm is proposed during the imputation process. Following theoretical discussions, the performance of the new methodology is evaluated through simulation studies and its application is demonstrated using an IPD set from patients with kidney failure.

C1278: **Capturing asymmetric structures and separability in multivariate contingency tables based on f-divergence**
Presenter: **Hisaya Okahara**, Tokyo University of Science, Japan
Co-authors: Kouji Tahata

A novel extension of asymmetry models is introduced for multivariate contingency tables with ordinal categories based on f-divergence. The proposed model generalizes existing asymmetry models while maintaining a focus on symmetric structures, offering a flexible approach to capturing complex dependence patterns. Theoretical properties of the model are established, extending known results for existing models. These include the decomposition of the symmetry model and the asymptotic properties of likelihood ratio statistics, reinforcing the natural extension of the proposed framework. By incorporating various divergence measures, this methodology provides a unified and adaptable approach for analyzing multivariate categorical data. The model's performance evaluation using real-world data, along with a comparison with conventional approaches in terms of goodness-of-fit, will be presented.

C1195: **Goodness of fit assessment of item response theory models for binary data**
Presenter: **Federico Bacchi**, University of Bologna, Italy
Co-authors: Maria Rosaria Ferrante, Pinuccia Pasqualina Calia

The goodness of fit assessment of item response theory (IRT) models can be performed based on three main underpinnings: (i) relying on a dichotomous decision strategy using chi-square tests; (ii) relying on indices that quantify the degree of fit along a continuum; and (iii) analyzing the eigenvalues of the manifest variables' polychoric correlation matrix. Despite a large stream of literature, there is currently no consensus on how to unequivocally establish whether an estimated model is sufficiently good for the data. A simulation study on large binary data sets with a common underlying generating process (major factor domain) was performed to address this gap. Three degrees of misspecifications were explored through a minor factor domain consisting of 50 orthogonal common factors accounting for three different fixed proportions of variance (0%, 10%, 30%). Furthermore, two different sample sizes and three different levels of correlation among the major factors were considered. The results suggested that chi-square tests are well-calibrated only in the medium correlation specification with no variance accounted for by minor factors, while the fit indices strongly depend on the covariance structure of the major factor domain. The rules of thumb based on the eigenvalues tend to converge as the sample size increases. However, parallel analysis seemed to be more effective than counting eigenvalues greater than one for detecting the number of major factors.

C1323: **Orthogonal decomposition of probability tables with Aitchison geometry for symmetry assessment**
Presenter: **Keita Nakamura**, Tokyo University of Science, Japan
Co-authors: Tomoyuki Nakagawa, Kouji Tahata

We propose a novel approach to analyzing symmetry in two-way contingency tables using Aitchison geometry of the simplex. We focus on square contingency tables with identical row and column classifications, introducing a method that identifies symmetric probability tables as a linear subspace within the Aitchison geometry. This enables orthogonal projection of any probability table onto the symmetric subspace, yielding the nearest symmetric table. A key feature of this projection is that each cell in the resulting symmetric table is represented by the normalized geometric mean of corresponding symmetric cells in the original table. This characterization differs from standard maximum likelihood estimators, except when the original table is already symmetric. The probability tables are decomposed into symmetric and skew-symmetric components, which are orthogonal to each other. We develop measures to quantify the degree of asymmetry and present a symmetry test based on these measures using a parametric bootstrap approach. The practical application of the method is demonstrated through an example using real-world data.

CC455 Room BH (SE) 1.05 FORECASTING I Chair: Michael Owyang

C0835: **A framework for analyzing the cost-benefit tradeoff of training neural networks**
Presenter: **Simon Spavound**, LeBow College of Business Drexel University, United States
Co-authors: Nikolaos Kourentzes

Recent computational advances, along with interest in applications such as large language models, have led to the deployment of larger and larger deep neural networks. Irrespective of network size, their training is complex and stochastic and affects the quality of the model outputs. Achieving reliable results may require multiple training runs, which substantially increases computational costs. Moreover, large models can be computationally intensive even after training. One socially disadvantageous side effect is the cost, both monetary and environmental, of the training of such models. Little practical advice is available, which demonstrates the tradeoff between training models multiple times for increased accuracy or reliability vs the computational cost of such training. This is illustrated on a time series forecasting experimental setup to provide some guidance in this area and open up future avenues of exploration. The tradeoffs are demonstrated for local as well as global models and between them.

C1334: **Testing and quantifying economic resilience**
Presenter: **Yohei Yamamoto**, Hitotsubashi University, Japan

A formal testing procedure is proposed to examine the resilience of an economy. The approach remains applicable even when a cross-section of the control group is unavailable and circumvents potential bias in time-series regressions using data that includes structural breaks. Measures of shock absorption and cumulative recovery are provided. Empirical analysis reveals that most of the advanced countries were not resilient to the global financial crisis, while many were so during the COVID-19 pandemic. Potential determinants of economic resilience, such as financial leverage and labor market regulation, may have negative correlations with these measures, and other determinants have heterogenous associations depending on the nature of the crisis.

C1414: **Voting-based ex ante method for selecting strategy of the price characteristics prediction on real estate market**
Presenter: **Alicja Wolny-Dominiak**, University of Economics in Katowice, Poland
Co-authors: Tomasz Zadlo, Adam Chwila, Monika Hadas-Dyduch, Tomasz Stachurski, Malgorzata Krzciuk

The topic of selecting an optimal prediction strategy is addressed when utilizing parametric or nonparametric regression models. The term "prediction strategy" is understood as the pair: the assumed model and the prediction algorithm. It emphasizes the significance of ex-ante prediction accuracy, ensemble methodologies, and forecasting not only the values of the dependent variable but also a function of these values, such as the total or median. The research proposes a methodology for selecting a strategy to predict the vector of functions of the dependent variable using various ex-ante accuracy measures. The final decision is determined through a voting mechanism, wherein the candidates are prediction strategies and the voters are diverse prediction models with their respective prediction errors. As the method is based on a Monte Carlo simulation, it facilitates the consideration of novel scenarios not previously observed. First, a comprehensive theoretical description of the proposed method is provided, while subsequently, its practical application in predicting characteristics of prices on the real estate market is presented. The empirical example utilizes data from the USA market. All computational analyses are conducted using the R programming environment.

C1561: **Similarity-based path forecasting of U.S. recession periods**
Presenter: **Henri Nyberg**, University of Turku, Finland
Co-authors: Visa Kuntze, Samuel Rauhala

A nonparametric similarity-based approach is developed to obtain path forecasts for binary time series by finding probability forecasts for each viable sequence of observations multiple periods ahead. In contrast to the common way of specifying forecast horizon-specific parametric models, the path forecasts are internally consistent and obtained simultaneously for all the horizons. In an empirical illustration, the state of the U.S. business cycle is forecasted around the past three recession periods.

CC500 Room BH (SE) 2.10 MACHINE LEARNING IN ECONOMICS AND FINANCE II Chair: Yoosoon Chang

C0636: **Calibrating option pricing models using neural networks and population-based optimization methods**
Presenter: **Antonio Santos**, University of Coimbra, Portugal
Co-authors: Jose Luis Esteves dos Santos, Margarida Biscaia Caleiras

In finance, model calibration is a crucial task that ensures that financial models accurately reflect market conditions, reducing the risk of decisions based on unreliable information. However, this calibration process is often computationally intensive and time-consuming, especially when dealing with complex models. To address these challenges, neural networks have emerged as a promising approach for developing more efficient option pricing methods, consequently allowing the utilization of algorithms that enhance the calibration process. The aim is to present a novel implementation and comparative analysis of the performance of two types of neural networks, feedforward neural networks (FNN) and long short-term memory networks (LSTM), in solving the Heston model calibration problem. A two-step calibration approach is employed, using neural networks to approximate the pricing function and significantly reduce calibration time. Numerical experiments demonstrate that LSTM networks, particularly when combined with a convergent variant of the differential evolution algorithm, can improve calibration accuracy compared to FNNs.

C1539: **Machine learning applications for the valuation of options on non-liquid option markets**
Presenter: **Jiri Witzany**, University of Economics in Prague, Czech Republic
Co-authors: Milan Ficura

Recently, there has been considerable interest in machine learning (ML) applications for the valuation of options. The main motivation is the speed of calibration or, for example, the calculation of credit valuation adjustment. It is usually assumed that there is a relatively liquid market with plain vanilla option quotations that can be used to calibrate the volatility surface or to estimate the parameters of an advanced stochastic model. In the second stage, the calibrated volatility surface is used to value given exotic options, again using a trained neural network (NN). The NNs are typically trained offline by sampling many models and market parameter combinations and calculating the options market values. The focus is on the quite common situation of a nonliquid option market in which one lacks a sufficient number of plain vanilla option quotations to calibrate the volatility surface, but one still needs to value an exotic option or just a plain vanilla option, subject to a more advanced stochastic model, as is typical for energy and carbon derivatives markets. It is shown that the historical return moment-based pricing and calibration NNs can be applicable in practice with a performance lower than but still comparable to the option-based calibration. It is also demonstrated that the performance can be substantially improved when high-frequency historical data, allowing them to apply the concept of realized volatility, are available.

C1553: **Machine learning for commodity futures pricing**
Presenter: **Milan Ficura**, University of Economics in Prague, Czech Republic

The ability of 29 characteristics is tested, previously reported in the literature, to predict the cross-section of 28 US commodity futures returns in the period from March 1996 to January 2024. The characteristics include momentum, volatility, liquidity, basis, relative basis, basis momentum, hedging pressure, speculative crowding, currency betas, inflation betas, stock market betas and commodity market betas, among others. In addition to the standard methodology based on univariate and multi-variate portfolio sorts, the ability of machine-learning methods (Elastic net regression and random forests) is further tested to extract predictive signals from the vectors of individual characteristics and predict the cross-section of commodity futures returns. The constructed models are found to possess significant out-of-sample predictive power and result in sizable economic gains.

C1652: **Prediction of local extrema in financial time series with multiple timeframe extreme gradient boosting method**
Presenter: **Chun Fai Carlin Chu**, The Hang Seng University of Hong Kong, Hong Kong
Co-authors: Po Kin David Chan

The prediction of maximum and minimum values in a financial time frame is crucial for developing automatic trading strategies. The existence of noise and non-stationarity inherited in the series challenges the effectiveness of traditional time series methods, and recent literature has demonstrated that the use of machine learning methods with richer input features is more promising. The financial time series is first processed by smoothing functions to alleviate the influence from random noise. The resultant series is processed for the identification of a set of pseudo local extrema, and their locations are subsequently modelled by multiple Bayesian optimized XGBoost models using a set of Shapley additive explanations

(SHAP) selected features derived from level 2 data. The models are tuned to achieve maximum precisions, minimizing the chance of unfavourable trades. Afterwards, the extrema prediction model is integrated with a trading algorithm for the development of customer-centric strategies. The proposed method considers extrema prediction, market condition and customer risk collectively. It is evaluated empirically with 15-day real market data, and its VWAP outperforms several benchmarks. On the other hand, an effective data handling procedure to consolidate irregularly observed level 1 and level 2 data with the consideration of sampling bias will be addressed in this research work.

Monday 16.12.2024 14:40 - 16:20 Parallel Session P – CFECMStatistics2024

CO272 Room S-2.25 Semi-parametric inference via data integration (virtual) Chair: Abhishek Chakrabortty

C1428: **Robust and efficient high-dimensional inference with surrogate outcomes**
Presenter: **Yong Chen**, Univ. of Pennsylvania, United States

Electronic health records (EHR) offer a valuable resource for discovering novel disease risk factors. However, the common issue of missingness in the primary phenotype of interest often leads to efficiency loss in inferential methods that rely solely on fully observed samples. Additionally, the prevalent misclassification of EHR-derived phenotypes can result in systematic bias, thereby affecting the reproducibility of findings. In response to these challenges, a robust and efficient framework for high-dimensional EHR-based discovery is introduced. Based on a class of surrogate models for EHR-based phenotypes, an augmented score function is constructed, and a corresponding test statistic is developed. The statistic not only maintains correct coverage under the null hypothesis but also exhibits enhanced power under local alternatives, outperforming tests that only use fully observed samples. Surprisingly, it achieves the correct coverage even in scenarios with arbitrary misclassification of EHR-based phenotypes and misspecified surrogate models. The statistical effectiveness of the proposed method is evaluated through extensive simulations and real-world data application.

C0231: **Efficient federated learning of the average treatment effect**
Presenter: **Sijia Li**, Harvard University, United States

A new data fusion method is introduced that utilizes multiple data sources to estimate the average treatment effect. Most existing methods only make use of fully aligned data sources that share common conditional distributions of the outcome. However, in many settings, the scarcity of fully aligned sources can make existing methods require unduly large sample sizes to be useful. The approach enables the incorporation of weakly aligned data sources that are not perfectly aligned, provided their degree of misalignment is known up to finite-dimensional parameters. The canonical gradient is derived to estimate the average treatment effect in such a data fusion setting, and the semiparametric efficiency bound is characterized. Furthermore, the proposed approach is decentralized and only requires individual-level data from the user-specified target data and summary-level statistics from other sources.

C0261: **Continuous treatment effects with surrogate outcomes**
Presenter: **Zhenghao Zeng**, Carnegie Mellon University, United States
Co-authors: David Arbour, Avi Feller, Raghavendra Addanki, Ryan Rossi, Ritwik Sinha, Edward Kennedy

In many real-world causal inference applications, the primary outcomes (labels) are often partially missing, especially if they are expensive or difficult to collect. If the missingness depends on covariates (i.e., missingness is not completely at random), analyses based on fully observed samples alone may be biased. Incorporating surrogates, which are fully observed post-treatment variables related to the primary outcome, can improve estimation in this case. The role of surrogates is studied in estimating continuous treatment effects and propose a doubly robust method to efficiently incorporate surrogates in the analysis, which uses both labeled and unlabeled data and does not suffer from the above selection bias problem. Importantly, the asymptotic normality of the proposed estimator is established, and possible improvements in the variance are shown compared with methods that solely use labeled data. Extensive simulations show the methods enjoy appealing empirical performance.

C0721: **Enhancing efficiency and robustness in high-dimensional linear regression with additional unlabeled data**
Presenter: **Yuqian Zhang**, Renmin University of China, China
Co-authors: Kai Chen

In semi-supervised learning, the prevailing understanding suggests that observing additional unlabeled samples improves estimation accuracy for linear parameters only in the case of model misspecification. The aim is to challenge this notion, demonstrating its inaccuracy in high dimensions. Initially focusing on a dense scenario, robust semi-supervised estimators are introduced for the regression coefficient without relying on sparse structures in the population slope. Even when the true underlying model is linear, it is shown that leveraging information from large-scale unlabeled data improves both estimation accuracy and inference robustness. Moreover, semi-supervised methods are proposed to further enhance efficiency in scenarios with a sparse linear slope. Diverging from the standard semi-supervised literature, covariate shifts are also allowed. The performance of the proposed methods is illustrated through extensive numerical studies, including simulations and a real-data application to the AIDS Clinical Trials Group Protocol 175 (ACTG175).

CO214 Room S-1.01 Statistical modelling of text data Chair: Bettina Gruen

C0320: **Time-varying Poisson factorization with an application to U.S. Senate speeches**
Presenter: **Jan Vavra**, PLUS, Austria
Co-authors: Bettina Gruen, Paul Hofmarcher

The world is evolving, and so is the vocabulary used to discuss various topics. However, current dynamic Poisson factorization models are designed for repeated observations on the same units, yielding matrices of counts for each considered time period. Such a format is not suitable for text-modelling purposes as documents usually cover very few topics, and their aggregation complicates the identifiability of the topics. Therefore, a time-varying Poisson factorization (TVPF) model is proposed that works with documents as the smallest observable units and captures the evolution of the popularity of words for each topic separately. The posterior distribution is approximated with variational inference, where the use of the mean-field variational family is questioned, especially for the time-varying components. Moreover, the frequently used random walk scheme is relaxed to a general AR(1) process. The stability and similarity of the topics are explored via a dissimilarity measure derived from the Kullback-Leibler divergence between variational families. The use of TVPF is illustrated in speeches from 18 sessions in the U.S. Senate, 1981-2016, where the primary focus is on the evolution of the climate change topic. Yet, for this example, both provided variational information criteria, and the simplicity of random walk parameterization estimated by mean-field variational inference is still preferred.

C0402: **One sample, one label: Learning from labels with different degrees of informativeness**
Presenter: **Matthias Assenmacher**, LMU Munich, Germany

Unlike classical tabular data commonly used in statistics, data sets utilized for text mining or natural language processing (NLP) can exhibit a wildly different structure, and so do the labels. For most tasks of interest to NLP research, it is not as easy as just measuring the values of the target variable; instead, human labour often has to be employed for this purpose. This can have several implications when developing and training models: Human annotations of text can be highly subjective (depending on the task and the data), they might incur different costs, or they could altogether be challenging to come up with, as the existence of only one ground truth gold label itself is highly debatable. The probably most prominent example of the latter is the case of open-ended text generation, a task in which the model's capabilities have recently made a significant leap upon the introduction of large language models (LLM). However, the challenge of evaluating the LLM-generated text is still far from being solved due to subjective criteria for the desired outputs. A further example is topic modeling, where the target itself (i.e. the topic distribution in a document) is a latent variable to be estimated before it can be associated with additional covariates.

C0475: **Monitoring (social) media narratives combining retrospective few-shot classification with continuous topic modeling**
Presenter: **Jonas Rieger**, TU Dortmund University, Germany

A variety of topic models and text classification models exist, each with distinct characteristics. However, in truly interdisciplinary applications, many of such models prove unsuitable due to specific individual limitations such as continuously growing text corpora or (huge) unbalancedness of classification labels. These issues are addressed by meshing suitable supervised and un-/semi-supervised learning approaches to efficiently visualize thematic trends in texts in real-time using dynamic topic models while enabling diachronic quantification of argumentative shifts and their uncertainties using static few-shot learning techniques to enable parameter-efficient fine-tuning of transformer-based pre-trained language models. This hybrid approach facilitates rapid insights into thematic trends and allows for retrospective quantification with the need for only small-scale manual coding experiments. Presenting different performance and reliability measures, straightforward usability and state-of-the-art performances are demonstrated for argumentative example datasets covering different thematic areas.

C0818: **Seeded Poisson factorization: Leveraging domain knowledge to fit topic models**
Presenter: **Bernd Prostmaier**, BMW AG, Germany
Co-authors: Bettina Gruen, Paul Hofmarcher

The latent variable model seeded Poisson factorization (SPF) is proposed, which addresses the challenges in text classification where no labelled texts are available, but the classes are characterized with a set of relevant words. In various business contexts, including, in particular, the assessment of consumer feedback, vast amounts of unlabeled text data are collected where conceptual frameworks outline potential categorization schemata, and domain experts are able to provide sets of relevant words for each category. SPF builds on the Poisson factorization topic model which assumes that term counts in documents are independently drawn from a Poisson distribution with the rate resulting from a combination of topic-specific term distributions weighted by the document-specific topic distributions. Seeding modifies the prior distribution of the topic-specific term distributions with the set of relevant words a-priori having higher rates for their topic. Estimation is based on computationally efficient variational inference using general-purpose stochastic gradient optimization. The use of SPF is illustrated on Amazon customer feedback data to classify feedback items where the categories are a-priori known. Empirical results indicate that SPF surpasses alternative topic models, allowing for the specification of seed words for topics in terms of computational cost and classification accuracy.

CO003 Room S-1.06 MACHINE LEARNING AND OPTIMIZATION WITH APPLICATIONS Chair: Chengchun Shi

C0213: **Harnessing geometric signatures in causal representation learning**
Presenter: **Yixin Wang**, University of Michigan, United States

Causal representation learning aims to extract high-level latent causal factors from low-level sensory data. Many existing methods often identify these factors by assuming they are statistically independent. In practice, however, the factors are often correlated, causally connected, or arbitrarily dependent. The purpose is to explore how one might identify such dependent causal latent factors from data, whether passive observations, interventional experiments, or multi-domain datasets. The key observation is that, despite correlations, the causal connections (or the lack of) among factors leave geometric signatures in the latent factors' support -the ranges of values each can take. Leveraging these signatures, it is shown that observational data alone can identify the latent factors up to coordinate transformations if they bear no causal links. When causal connections do exist, interventional data can provide geometric clues sufficient for identification. In the most general case of arbitrary dependencies, multi-domain data can separate stable factors from unstable ones. Taken together, these results showcase the unique power of geometric signatures in causal representation learning.

C0460: **Bi-level offline reinforcement learning**
Presenter: **Wenzhuo Zhou**, University of California Irvine, United States

Offline reinforcement learning (RL) is studied to learn a good policy based on a fixed, pre-collected dataset. A fundamental challenge behind this task is the distributional shift due to the dataset lacking sufficient exploration, especially under function approximation. To tackle this issue, a bi-level structured policy optimization algorithm is proposed that models a hierarchical interaction between the policy (upper level) and the value function (lower level). The lower level focuses on constructing a confidence set of value estimates that maintain sufficiently small weighted average Bellman errors while controlling uncertainty arising from distribution mismatch. Subsequently, at the upper level, the policy aims to maximize a conservative value estimate from the confidence set formed at the lower level. This novel formulation preserves the maximum flexibility of the implicitly induced exploratory data distribution, enabling the power of model extrapolation. In practice, it can be solved through a computationally efficient, penalized adversarial estimation procedure. The theoretical regret guarantees do not rely on any data-coverage and completeness-type assumptions, only requiring realizability. These guarantees also demonstrate that the learned policy represents the best effort among all policies, as no other policies can outperform it.

C0608: **Local optimization for sequential design of experiments via sparse meta-model**
Presenter: **Matteo Borrotti**, University of Milan-Bicocca, Italy
Co-authors: Davide Ferrari

High-dimensional experiments can be characterized by many input variables and, often, a limited number of observations. The final aim is to optimize the experimental response. A possible solution is local optimization, suitable for expensive, high-dimensional black-box experiments. One possible advantage of local optimization is that it does not need to explore the entire experimental space to reach the optimum. The proposal is to use a local probabilistic model to estimate the objective response surface. The local meta-model is used to guide the search on the objective response surface and to provide an automatic way of determining the step size from one iteration to the next. Furthermore, to handle a possible issue related to sparsity on the input variables, the proposed approach detects the most important variables at each step to move in significant directions of the objective response surface. The final solution is compared with a benchmark on different simulation data models.

C0949: **Doubly robust interval estimation for optimal policy evaluation in online learning**
Presenter: **Hengrui Cai**, University of California Irvine, United States

Evaluating the performance of an ongoing policy plays a vital role in many areas, such as medicine and economics, to provide crucial instruction on the early stop of the online experiment and timely feedback from the environment. Policy evaluation in online learning thus attracts increasing attention by inferring the mean outcome of the optimal policy (i.e., the value) in real time. Yet, such a problem is particularly challenging due to the dependent data generated in the online environment, the unknown optimal policy, and the complex exploration and exploitation trade-off in the adaptive experiment. The aim is to overcome these difficulties in policy evaluation for online learning. The probability of exploration that quantifies the probability of exploring the non-optimal actions is explicitly derived under commonly used bandit algorithms. This probability of conducting valid inference is used on the online conditional mean estimator under each action, and the doubly robust interval estimation (DREAM) method is developed to infer the value under the estimated optimal policy in online learning. The proposed value estimator provides double protection on the consistency and is asymptotically normal with a Wald-type confidence interval provided. Extensive simulations and real data applications are conducted to demonstrate the empirical validity of the proposed DREAM method.

CO160 Room S-1.27 ON KERNEL ESTIMATION ON MANIFOLDS, TESTING OUTLIERS AND MINIMAX RISK Chair: Anne Francoise Yao

C1203: **Minimax risk with random normalizing factors in the single-index model**
Presenter: **Armel Fabrice Evrard Yode**, Universite Felix Houphouet-Boigny, Cote d'Ivoire

The nonparametric estimation problem of a multidimensional regression function is considered. The aim is to propose improvement in the optimal estimation rate from the minimax point of view. To avoid poor estimation quality or generally in models for which the minimax approach is unsatisfactory, a prior study introduced the concept of random normalizing factors in 1999. This concept is a combination of adaptive estimation and minimax hypothesis testing theory. This hybrid approach uses the results of test theory to consider adaptive estimation. So, via the concept of random normalizing factors introduced by the prior study, considering a plausible assumption that the regression function has a single-index structure, an estimator that can be adaptive is constructed and whose observation-dependent estimation rate is better than that obtained via the minimax approach, with prescribed confidence level n. In addition, the relevance of the results is demonstrated by applying them to real data sets.

C1199: **Kernel density estimation for continuous Riemannian stochastic processes**
Presenter: **Anne Francoise Yao**, Universite Clermont Auvergne/LMBP, France
Co-authors: Vincent Monsan, Axel Mothe, Djack Guy-Aude Kouadio, Catherine Aaron

The focus is on kernel density estimation of the univariate marginal distribution of a strongly mixing continuous time process. This topic has been widely treated in the literature in the case where the process has values in an Euclidean space. However, the situation where the process lives in a Riemannian submanifold has not yet been treated. The estimator appears as the integral counterpart of a recent work, which generalized the results of a prior study to the case of stochastic processes under some mixing conditions. Namely, some consistent results and applications to some diffusion processes are given.

C1205: **Using EVT to test for outliers**
Presenter: **Nathaniel Gbenro**, Ensea-Abidjan, Cote d'Ivoire
Co-authors: Abdou Ka Diongue, El hadji Deme

The focus is on the identification of aberrant values by using extreme value theory (EVT) techniques. A new approach is proposed in the identification process of outliers. In this framework, the algorithm suggested by a prior study is extended to Gaussian or non-Gaussian distributions. Two empirical applications have been established to illustrate the efficiency of the approach. First, simulated data is used from Gaussian distributions, and the methodology is compared to the one proposed by a past study. In the second application, simulated data from various non-Gaussian distributions is also used to study the performance of our approach. The results suggest that the EVT outliers' test has good power when the sample size is large.

C1198: **Kernel regression estimation for stochastic process with values in a Riemannian manifold**
Presenter: **Mohamed Abdillahi Isman**, University Clermont Auvergne, France
Co-authors: Papa Mbaye, Salah Khardani, Anne Francoise Yao, Wiem Nefzi

The aim is to study the behavior of the kernel regression estimator when the output is a real-valued random variable, Y and the input, X, is a random variable which takes place in a finite-dimensional Riemannian submanifold. The results of a past study are extended to independent identically distributed observations in the case of dependent data under some mixing conditions. Specifically, the rate of convergence is given in mean square error meaning, probability, and almost surely. Furthermore, a central limit theorem is established, and the purpose is illustrated through some simulations and a real data application.

CO146 **Room Auditorium** **ECOSTA JOURNAL** **Chair: Cristian Gatu**

C0305: **Heterogeneous Graphon JSQ(d) model**
Presenter: **Arka Ghosh**, , United States
Co-authors: Yan-Han Chen, Ruoyu Wu

A variation of the supermarket model is considered in which a task arriving at a dispatcher is routed to one of its neighborhood servers based on the JSQ(d) strategy. Both heterogeneous dispatchers and servers are considered whose neighborhood relationships are described by a deterministic graphon. The evolution of the queue length for each server is described in the form of stochastic differential equations in which the interaction between servers exists. The law of large number results, both locally and globally, are established as the size of the system grows and the underlying graphons converge. The interacting system is proven to converge to an independent but heterogeneous system.

C1160: **On modified Anderson-Darling statistic for various distributions with unknown parameters**
Presenter: **Hidetoshi Murakami**, Tokyo University of Science, Japan
Co-authors: Hikaru Yamaguchi

For a long time, numerous goodness-of-fit test statistics have been considered and applied in many scientific fields. One of the most powerful statistics is the Anderson-Darling statistic, which is sensitive to discrepancies at the tails of the distribution rather than near the center. However, for example, hydrologists are interested in estimates of flood magnitudes for high return periods. In these cases, the curiosity is concentrated on the upper tail of the distribution. Then, The aim is to propose the generalized modified Anderson-Darling statistic that emphasizes the upper or lower tails of the distribution. The limiting distribution of the proposed statistic is estimated by using both theoretical approximation and simulation. Under the finite sample sizes, the distribution of the proposed statistic is estimated via a simulation study. Simulations are used to investigate the power of proposed statistics for various distributions with unknown parameters. The proposed statistic is illustrated by the analysis of real data.

C0358: **Matrix-valued factor model with time-varying main effects**
Presenter: **Zetai Cen**, London School of Economics and Political Science, United Kingdom
Co-authors: Clifford Lam

A new proposal for the matrix-valued time-varying main effects factor model (MEFM) is presented. MEFM is a generalization to the traditional matrix-valued factor model (FM). Rigorous definitions of MEFM and its identifications are given, with estimators proposed for the time-varying grand mean, row and column main effects, and the row and column factor loading matrices for the common component. Rates of convergence for different estimators are spelt out, with asymptotic normality shown. The core rank estimator for the common component is also proposed, with the consistency of the estimators presented. A crucial test is proposed to test if FM is sufficient against the alternative that MEFM is necessary, whose power is demonstrated by various simulation settings. The accuracy of our estimators is also demonstrated numerically in extended simulation experiments. A set of NYC taxi traffic data is analyzed, and the proposed test suggests that MEFM is indeed necessary to analyze the data against a traditional FM.

C1425: **Stochastic representation, efficient computation methods, and stochastic ordering in tree-structured Ising models**
Presenter: **Etienne Marceau**, Laval University, Canada

High-dimensional multivariate Bernoulli distributions are essential in the modeling of binary data in actuarial contexts. Tree-structured Ising models, a class of undirected graphical models for binary data, have been proven to be useful in a variety of applications in machine learning. The advantages of expressing tree-based Ising models are assessed via their mean parameterization rather than their commonly chosen canonical parameterization. This includes fixed marginal distributions, often convenient for dependence modeling, and the dispelling of the intractable normalizing constant otherwise plaguing Ising models. An analytic expression for the joint probability-generating function of mean-parameterized tree-structured Ising models is derived. The latter is used to build efficient computation methods for the sum of its constituent random variables. Similarly, an analytic expression is derived from their ordinary generating function of expected allocations, providing means for exact computations in the context of risk allocations.

CO399 Room K0.16 RECENT ADVANCEMENTS IN STATISTICAL NETWORK ANALYSIS AND BEYOND Chair: Weijing Tang

C0408: **Community detection with heterogeneous block covariance model**
Presenter: **Yunpeng Zhao**, Colorado State University, United States

Community detection is the task of clustering objects based on their pairwise relationships. Most of the model-based community detection methods, such as the stochastic block model and its variants, are designed for networks with binary (yes/no) edges. In many practical scenarios, edges often possess continuous weights, spanning positive and negative values, which reflect varying levels of connectivity. To address this challenge, the heterogeneous block covariance model (HBCM) is introduced, which defines a community structure within the covariance matrix where edges have signed and continuous weights. Furthermore, it takes into account the heterogeneity of objects when forming connections with other objects within a community. A novel variational expectation-maximization algorithm is proposed to estimate the group membership. The HBCM provides provable consistent estimates of memberships, and its promising performance is observed in numerical simulations with different setups. The model is applied to a yeast gene expression dataset to detect the gene clusters regulated by different transcript factors during the yeast cell cycle.

C0846: **Efficient analysis of latent spaces in heterogeneous networks**
Presenter: **Yinqiu He**, University of Wisconsin - Madison, United States

Efficient estimation of the latent structures is studied for a collection of heterogeneous networks. A latent space model is proposed with a shared latent structure along with distinct individual structures. A procedure is developed that learns the shared space from the data. Estimation is achieved by parametric efficient score equations for the latent space parameters. Oracle error rates are derived to estimate both the shared and distinct latent space parameters simultaneously. The method and theory encompass a wide range of types of edge weights under exponential family distributions.

C0946: **Errorfully observed Markov models for evolving networks**
Presenter: **Peter MacDonald**, University of Waterloo, Canada
Co-authors: Eric Kolaczyk

A class of continuous-time Markov chain models is considered for binary network data, which evolves over time on an aligned set of nodes. Continuous and discrete-time observation schemes are investigated, where under discrete observation, inference is based on partial observation of the underlying continuous-time process. The statistical properties of some commonly used dynamic network summaries (edge density, number of grown or dissolved edges) are studied towards estimation and inference.

C1594: **Latent space directed counting network models and their application to citation networks of statistical journals**
Presenter: **Ji Zhu**, University of Michigan, United States

Impact factors evaluate a journal's significance. Specifically, for a given year, it is defined as the ratio between the number of citations received this year by publications in this journal in two preceding years and the number of publications in this journal in two preceding years. Although popular and straightforward, impact factors have two limitations. First, they do not distinguish between citations from different domains, counting all citations for a journal equally. Second, impact factors are one-dimensional metrics that overlook mutual citation information between journals, a two-dimensional data resource, which could provide valuable insights. A latent space directed counting network model is introduced, that explores latent variables driving the citation pattern among journals in the same domain through the analysis of mutual citation counts. The proposed model simultaneously takes into account a journal's general significance and its significance through interactions with other journals in the domain. Likelihood-based estimators of the parameters with their statistical optimality established are introduced. A simulation study verifies the theory and shows the effectiveness of the algorithm. A dataset consisting of mutual citation information among 104 statistical journals is collected and cleaned. Fitting the data with the latent space directed counting network model, meaningful and interpretable findings are uncovered that are not conveyed by impact factors.

CO169 Room K0.18 DEVELOPMENTS IN SPATIO-TEMPORAL MODELING OF HEALTH OUTCOME DATA Chair: Andrew Lawson

C0421: **Dengue nowcasting in Brazil by combining official surveillance data and Google Trends information**
Presenter: **Paula Moraga**, King Abdullah University of Science and Technology (KAUST), Saudi Arabia

Dengue is a mosquito-borne viral disease that poses significant public health challenges in tropical and sub-tropical regions worldwide. Effective surveillance systems are essential for dengue prevention and control. However, traditional systems rely on delayed data, limiting their effectiveness. The value of using Google Trends data to complement official dengue data is evaluated for nowcasting dengue in Brazil, a country frequently affected by this disease. Various nowcasting approaches are compared that incorporate autoregressive features from official dengue cases, Google Trends data, and a combination of both, using a naive approach as a baseline. The performance of these methods is evaluated by nowcasting weekly dengue cases from March to June 2024 across Brazilian states. Error measures and 95% coverage probabilities reveal that models incorporating Google Trends data enhance the accuracy of weekly nowcasts across states and offer additional insights into dengue activity levels. To support real-time decision-making, 'Dengue-Tracker' is also presented, a website that displays weekly nowcasts to inform both decision-makers and the public, improving situational awareness of dengue activity. In conclusion, the value of digital data sources in enhancing dengue nowcasting is demonstrated, and the importance of integrating alternative data streams into traditional surveillance systems is emphasized for better-informed decision-making.

C0706: **Latent archetypes of the spatial patterns of cancer**
Presenter: **Marcos Prates**, Universidade Federal de Minas Gerais, Brazil
Co-authors: Thais Pacheco, Renato Assuncao

The cancer atlas edited by several countries is the main resource for the analysis of the geographic variation of cancer risk. Correlating the observed spatial patterns with known or hypothesized risk factors is time-consuming work for epidemiologists who need to deal with each cancer separately, breaking down the patterns according to sex and race. The recent literature has proposed studying more than one cancer simultaneously while looking for common spatial risk factors. However, this previous study has two constraints: they consider only a very small (2-4) number of cancers previously known to share risk factors. An exploratory method is proposed to search for latent spatial risk factors of a large number of supposedly unrelated cancers. The method is based on the singular value decomposition and non-negative matrix factorization; it is computationally efficient, scaling easily with the number of regions and cancers. A simulation study is carried out to evaluate the method's performance, and it is applied to cancer atlas from the USA, England, France, Australia, Spain, and Brazil. It is concluded that with very few latent maps, which can represent a reduction of up to 90% of atlas maps, most of the spatial variability is conserved. The hope is that by concentrating on the epidemiological analysis of these few latent maps, a substantial amount of work is saved, and, at the same time, high-level explanations affecting many cancers can be simultaneously reached.

C0603: **Using wastewater data for COVID-19 surveillance in the post-pandemic era: A data integration approach**
Presenter: **Guangquan Li**, Northumbria University, United Kingdom
Co-authors: Peter Diggle, Marta Blangiardo

During the COVID-19 pandemic, wastewater-based epidemiology (WBE), a suite of methods to detect and measure viral contents in wastewater, has been recognised as an efficient surveillance tool to monitor the disease. WBE is used in the post-pandemic setting, where data collection via national randomised surveys is run at a reduced scale, but wastewater data, a spatially refined and low-cost metric, can be used to complement the reduced health data for cost-effective disease monitoring. Using data collected from a network of sewage treatment works, a geostatistical model is constructed to predict wastewater viral load at a fine space-time scale for the whole of England. A data integration framework is developed to combine these viral predictions with prevalence, estimated using data collected through randomised surveys and community testing. The data integration framework aims to produce prevalence nowcast at a fine spatial scale when prevalence estimates can only be derived at a coarse spatial level due to the scaled-down health data collection. The results from the cross-validation demonstrate the added values of wastewater data, not only improving the accuracy of the prevalence nowcast but also reducing the nowcast uncertainty. The investigation also highlights the critical role of the coarse-level prevalence estimates in anchoring the wastewater data, thus calling for the need to maintain some form of reduced-scale national prevalence survey in the non-pandemic periods.

C0720: **A novel Bayesian spatiotemporal surveillance metric to predict emerging infectious disease high-risk clusters**
Presenter: **Joanne Kim**, The Ohio State University, United States
Co-authors: Andrew Lawson

Identification of high-risk disease clusters has been one of the top goals of infectious disease public health surveillance. However, previous spatial disease mapping research focused on identifying the current hotspot of the elevated risk area. Still, it did not provide information about where the next high-risk cluster is likely to occur, given the existing hotspot. A novel Bayesian metric is introduced to predict the occurrence of new clusters of the elevated risk areas for the infectious disease outbreak. The proposed metric utilizes the areas' own risk profile, temporal risk trend, and spatial neighborhood influence. A weighting scheme is also introduced to balance these three components, which accommodates the characteristics of the infectious disease outbreak, and spatial disease trends. Thorough simulation studies were conducted to identify the optimal weighting scheme and evaluate the performance of the proposed cluster prediction surveillance metric. Results indicate that the areas' own risk and the neighborhood influence play an important role in making a highly sensitive metric, and the risk trend term is important for the specificity and the accuracy of prediction.

CO082 Room K0.19 CAUSAL INFERENCE Chair: Massimo Cannas

C0342: **A continuous-time joint marginal structural survival model for causal inferences about multiple intermittent treatments**
Presenter: **Liangyuan Hu**, Rutgers University, United States
Co-authors: Himanshu Joshi, Erick Scott , Fan Li

To draw real-world evidence about the comparative effectiveness of multiple time-varying treatments on patient survival, a joint marginal structural survival model and a novel weighting strategy are developed to account for time-varying confounding and censoring. The methods formulate complex longitudinal treatments with multiple start/stop switches as the recurrent events with discontinuous intervals of treatment eligibility. The weights are derived in continuous time to handle a complex longitudinal dataset without the need to discretize or artificially align the measurement times. Machine learning models, designed for censored survival data with time-varying covariates and the kernel function estimator of the baseline intensity, are further used to efficiently estimate the continuous-time weights. Simulations demonstrate that the proposed methods provide better bias reduction and nominal coverage probability when analyzing observational longitudinal survival data with irregularly spaced time intervals compared to conventional methods that require aligned measurement time points. The proposed methods are applied to a large-scale COVID-19 dataset to estimate the causal effects of several COVID-19 treatments on the composite of in-hospital mortality and ICU admission

C0571: **The comparison of MARMoT adjustment and template matching in a multiple treatment framework: A simulation study**
Presenter: **Margherita Silan**, Department of Statistical Sciences, University of Padova, Italy
Co-authors: Pietro Belloni

Specific statistical tools are required to estimate a causal effect when many treatments are involved. Case studies involving such a number of treatments are rare in the scientific literature and typically are based on two main methods: Matching on poset-based average rank for multiple treatments (MARMoT) and template matching. The aim is to compare the two techniques through a simulation study in various scenarios. Those artificial scenarios are built that vary in multiple aspects, such as the number of treatments (50, 250, 500), the presence of rare treatments, and the presence of rare confounders. In addition, minor technical adjustments were made to enhance the performance of both techniques in the different settings. The objective is to empirically determine which technique performs better in each scenario. These methods were also applied to real data from the Medicare database, comparing 41 medical facilities on their performance with elderly patients undergoing cardiac surgeries. Conclusions could provide valuable insights for choosing and implementing statistical methods to address self-selection bias in multi-treatment observational studies.

C0586: **Bounds and identification for the probability of causation in individual cases**
Presenter: **Giuseppe Demuru**, Universita di Cagliari, Italy
Co-authors: Monica Musio, Philip Dawid

The purpose is to consider the problem of assessing whether, in an individual case, there is a causal relationship between an observed exposure and a response variable. When data are available on similar individuals, we may be able to estimate prospective probabilities, but even under ideal conditions, these are typically inadequate to identify the "probability of causation". Instead, bounds can be only derived for this. These bounds can be improved or amended when further knowledge is available. Current literature includes knowledge on pre-treatment covariates, unobserved confounding and mediators. New bounds are proposed for the case in which instrumental variables and/or observational data are available.

C0611: **Propensity score matching for cross-classified data structures**
Presenter: **Bruno Arpino**, University of Padua, Italy
Co-authors: Daniela Bellani

Cross-classified data structures arise in various settings where individual units can be grouped along two or more dimensions that are not nested within one another. As a motivating example, the effect of parenting style on the educational performance of children of immigrants is considered. When studying immigrants, it is common to account for the effects of both the country of origin and the country of destination, which together define a cross-classified structure. Additionally, it is often crucial to account for the community effect, which refers to the impact of the specific immigrant group residing in a particular destination country. Propensity score methods have been developed to address multilevel structures of different types. An approach is developed to match treated and control units within communities as much as possible. If this is not feasible, the method seeks matches within the country of destination or the country of origin, with a preference order that can be adjusted by the researcher on a case-by-case basis. The proposed approach's ability to balance confounders is evaluated, and the bias of causal estimates is reduced using simulated data, which is applied to data from the program for international student assessment (PISA).

CO313 Room K0.20 STATISTICAL INFERENCE AND ESTIMATION UNDER DEPENDENCE Chair: Rajarshi Mukherjee

C0653: **Statistical inference in Ising and Potts models**
Presenter: **Somabha Mukherjee**, National University of Singapore, Singapore

Some of the existing literature is summarized in the field of statistical inference in the classical Ising and Potts model, followed by presenting some of the recent results in the same area for tensor and more general regression versions of these classical models. The main focus is on consistent estimation of the inverse temperature and external field parameters of these models and proving asymptotics of these estimators. As seen, these asymptotics are accompanied by surprising phase transitions in terms of the true parameter position and the rate of convergence of the sufficient statistics. A comparative discussion is also given on the efficiencies of two of the most important consistent estimators in these models: the maximum likelihood and the maximum pseudolikelihood estimators.

C0588: **Classification under outcome misclassification: Reliability quantification and partial identification**
Presenter: **Muxuan Liang**, University of Florida, United States

Misclassification of outcomes or labels presents a prevalent challenge in classification problems. In many applications, the underlying outcome may not be directly accessible, while a surrogate outcome subject to misclassification can be observed. Directly using the surrogate outcome may lead to a biased estimation of the optimal classification rule. The sensitivity and specificity of the surrogate outcome at an individual level can be used to remove such bias. However, with limited accessibility of the underlying outcomes, point identification of individual-level sensitivity and specificity is difficult or even impossible. For classification problems, a range of individual-level sensitivity and specificity is assumed to be the reliable quantification of surrogate outcomes. With this partial information on sensitivity and specificity, partial identification is established for the distribution of the underlying outcome as well as the optimal classification rule using the surrogates. Based on this result, a robust classification framework and a novel estimation procedure are proposed to estimate a robust classification rule without requiring point identification of individual-level sensitivity and specificity.

C0645: **Detecting interference in A/B testing with increasing allocation**
Presenter: **Shuangning Li**, University of Chicago, United States
Co-authors: Kevin Han, Jialiang Mao, Han Wu

In the past decade, the technology industry has adopted A/B testing to guide product development and make business decisions. A/B tests are often implemented with increasing treatment allocation: the new treatment is gradually released to an increasing number of units through a sequence of randomized experiments. In scenarios such as experimenting in a social network setting or in a bipartite online marketplace, interference among units may exist, harming the validity of simple inference procedures. A procedure is introduced to test for interference in A/B testing with increasing allocation. The procedure can be implemented on an existing A/B testing platform with a separate flow and does not require a specific interference mechanism. In particular, two permutation tests that are valid under different assumptions are introduced. Firstly, a general statistical test is introduced for interference, requiring no additional assumption. Secondly, a testing procedure is introduced that is valid under a time-fixed effect assumption. The testing procedure has very low computational complexity, is powerful, and formalizes a heuristic algorithm already implemented in the industry. The performance of the proposed testing procedure is demonstrated through simulations on synthetic data. Finally, one application of the proposed methods at LinkedIn is discussed, where a screening step is implemented to detect potential interference in marketplace experiments.

C0800: **Statistical inference under dependent Gaussian mixture models**
Presenter: **Rajarshi Mukherjee**, Harvard T.H. Chan School of Public Health, United States

Gaussian mixture models are widely used to model data generated from multiple latent sources. Despite its popularity, most theoretical research assumes that the labels are independent and identically distributed, and it is unclear how the fundamental limits of estimation change under dependence. This question is addressed for the spherical two-component Gaussian mixture model with dependent labels. It is first shown that for labels with an arbitrary dependence, a naive estimator based on the misspecified likelihood is $\sqrt{n}$-consistent. Additionally, under labels that follow an Ising model, the information-theoretic limitations are established for estimation, and an interesting phase transition is discovered as dependence becomes stronger. The Ising model is a popular quadratic interaction model that allows network dependence. When the dependence is smaller than a threshold, the optimal estimator and its limiting variance exactly match the independent case. This result holds for a wide class of Ising models where the underlying network is dense enough. On the other hand, under stronger dependence, estimation becomes easier, and the naive estimator is no longer optimal. Hence, an alternative estimator is proposed based on the variational approximation of the likelihood, and its optimality is argued under a specific Ising model. In both cases, there is no information-computation gap, and the estimators are tractable.

CO299 Room K0.50 ADVANCES IN SUBSAMPLING AND ORDER-OF-ADDITION EXPERIMENTS (VIRTUAL) Chair: Nicholas Rios

C1122: **Subsampling in transfer learning**
Presenter: **Jing Wang**, University of Connecticut, United States

Transfer learning is an emerging field in recent years. Subsampling can be considered as a data selection method for transfer learning to obtain better performances. However, optimal subsampling with potential model misspecification has not been fully investigated, which limits the usage of subsampling algorithms in transfer learning. Subsampling algorithms are developed with potential mean shifts, which connects subsampling under misspecified models with data selection for transfer-learning algorithms. Theoretical analysis implies that the performances of transfer learning estimators are determined by model biases and variances. Therefore, two different subsampling strategies are proposed: one reduces model biases, and the other reduces variances due to subsampling. Two approaches are also proposed to combine the two sampling strategies to further improve the performances of transfer-learning estimators. Non-asymptotic bounds of the proposed estimators are proved. Numerical experiments justify the usage of the proposed transfer learning algorithms with data selection techniques.

C1080: **Sampling big data for model building using dimension reduction**
Presenter: **Ching-Chi Yang**, University of Memphis, United States

Handling the extraordinary data volume generated in many fields is challenging with current computational resources and techniques, especially when applying conventional statistical methods to big data. A common approach is to select sub-data that represent the full data. However, the sub-data should be carefully selected based on different objectives, such as rare event detection. Recent developments and newly published methods are introduced, and a broad discussion on potential research opportunities in the design of experiments is initiated.

C1579: **Modeling and designs for pairwise constrained order-of-addition experiments**
Presenter: **Xueru Zhang**, Purdue University, United States

In an order-of-addition (OofA) experiment, the sequence of m different components can significantly affect the response of the experiment. In many OofA experiments, the components are subject to constraints, where certain orders are impossible. For example, in survey design and job scheduling, the components are often arranged into groups, and these groups of components must be placed in a fixed order. If two components are in different groups, their pairwise order is determined by the fixed order of their groups. Design and analysis are needed for these pairwise-group constrained OofA experiments. A new model is proposed to accommodate pairwise-group constraints. A model is also introduced for mixed-pairwise constrained OofA experiments, which allows one pair of components within each group to have a pre-determined pairwise order. It is proven that the full design, which uses all feasible orders exactly once, is D- and G- optimal under the proposed models. Systematic construction methods are used to find optimal fractional designs for pairwise-group and mixed-pairwise constrained OofA experiments. The proposed methods efficiently assess the impact of question order in a survey dataset, where participants answered generalized intelligence questions in a randomly assigned order under mixed-pairwise constraints.

C1574: **Exact designs for OofA experiments under a transition-effect model**
Presenter: **Jiayi Zheng**, George Mason University, United States
Co-authors: Nicholas Rios

In the chemical, pharmaceutical, and food industries, sometimes the order of adding a set of components has an impact on the final product. These are instances of the order-of-addition (OofA) problem, which aims to find the optimal sequence of the components. Extensive research on this topic has been conducted, but almost all designs are found by optimizing the D-optimality criterion. However, when prediction of the response is important, there is still a need for I-optimal designs. A new model for OofA experiments is presented that uses transition effects to model the effect of order on the response, and the model is extended to cover cases where block-wise constraints are placed on the order of addition. Several algorithms are used to find both D and I efficient designs under this new model for many run sizes and for large numbers of components. Finally, two examples are shown to illustrate the effectiveness of the proposed designs and model in identifying the optimal order of addition, even under block-wise constraints.

CO031 Room K2.31 (Nash Lec. Theatre) FLEXIBLE LEARNING IN COMPLEX DATA ENVIRONMENTS Chair: David van Dyk

C1454: **Separating states in astronomical sources using hidden Markov models**

Presenter: **Robert Zimmerman**, University of Toronto, Canada
Co-authors: David van Dyk, Vinay Kashyap, Aneta Siemiginowska

A new method is presented to distinguish between different states (e.g., high and low, quiescent and flaring) in astronomical sources with count data. The method models the underlying physical process as latent variables following a continuous-space Markov chain that determines the expected Poisson counts in observed light curves in multiple passbands. Several autoregressive processes are considered for the underlying state process, yielding continuous-space hidden Markov models of varying complexity. Under these models, the state that the object is in can be inferred at any given time. The continuous state predictions from these models are then dichotomized with the help of a finite mixture model to produce state classifications. These techniques are applied to X-ray data from the active dMe flare star EVLac, splitting the data into quiescent and flaring states. It is found that a first-order vector autoregressive process efficiently separates flaring from quiescence: flaring occurs over 30-40% of the observation durations, a well-defined persistent quiescent state can be identified, and the flaring state is characterized by higher temperatures and emission measures.

C1583: **Experimental design for modern settings: Stories about text**

Presenter: **Alexander Volfovsky**, Duke University, United States

Given two texts, the question is which one is more persuasive. Such a comparison only informs about these two texts and does not inform what elements of the text drive the causal mechanism. Since the mechanism is of interest, a tempting design is to show many texts, measure their effects, and use natural language processing to learn what features of the texts should be considered as components of a causal analysis. However, such a black-box approach (e.g. a large language model) provides insufficient control of the causal model and may lead to spurious or nonsensical results. The question of what the treatment is is first addressed when the aim is to experiment with text. Specifically, the necessary (usually unstated) assumptions are outlined to make the text a plausible treatment. A novel experimental design is then developed that allows the researcher to control which elements of the text are being studied. For example, text is generated to study the effect of using intellectually humble language on the persuasiveness of the underlying text. Two major issues with machine learning methods are found for inferring causal effects of text. Transformer models that use learned representations of text as confounders overfit the data, inducing positivity violations. Other estimators that try to correct for text indirectly underfit the data and act like estimators that never even looked at text confounders.

C1436: **Model comparison for Bayesian lasso-like regression**

Presenter: **Christopher Hans**, The Ohio State University, United States
Co-authors: Ningyi Liu

Formal Bayesian methods for model comparison depend on the marginal likelihood of the data given a model. When closed-form expressions for marginal likelihoods in a given model class are not available, it is common to employ computational approaches to either estimate the marginal likelihoods directly or to avoid their explicit evaluation by summarizing output from carefully constructed MCMC algorithms. While Bayesian treatments of approaches to penalized regression have become popular tools for data analysis, formal Bayesian model comparison in these settings can be challenging. Computing marginal likelihoods for Bayesian lasso-like regression models is difficult due to the L1-norm penalty term that is incorporated into the prior on the regression coefficients. MCMC-based approaches are introduced for estimating the marginal likelihoods for the Bayesian lasso and elastic net regression models. The methods involve sampling from standard probability distributions, need not rely on any data augmentation, and require no tuning of random walks or specification of approximating distributions. Comparisons are made to other related approaches for marginal likelihood estimation.

C1505: **Doubly robust pivotal pointwise confidence intervals for a monotonic continuous treatment effect curve**

Presenter: **Charles Doss**, University of Minnesota, United States

A large majority of literature on evaluating the significance of a treatment effect based on observational data has been focused on discrete treatments. These methods are not applicable to drawing inferences for a continuous treatment, which arises in many important applications. Doubly robust confidence intervals are developed for the continuous treatment effect curve (at a fixed point) under the assumption that it is monotonic by developing a likelihood ratio-type procedure. Monotonicity is often a very natural assumption in the setting of a continuous treatment effect curve, and the assumption of monotonicity removes the need to choose a smoothing parameter for the nonparametrically estimated curve (or the related need to estimate the curve's unknown bias, which is challenging). The new methods are illustrated via simulations and a study of a dataset relating the effect of nurse staffing hours on hospital performance.

CO223 Room K2.40 RECENT DEVELOPMENTS IN NON-EUCLIDEAN STATISTICS Chair: Andrew Wood

C0391: **To the limit and beyond for the frequency modulated Mobius periodic regression model**

Presenter: **John Kent**, University of Leeds, United Kingdom
Co-authors: Charles C Taylor, Norah Almasoud

The frequency-modulated Mobius periodic regression model is an elegant and tractable model to describe how a real-valued response depends on a periodic explanatory variable, e.g. time with a daily cycle. The underlying idea is to warp time using a Mobius transformation and then to model the expected response as a first-order Fourier function of the warped time. Although the model is simple to describe and straightforward to fit for "nice" data, there are various subtle features worthy of deeper study. First, two summary measures are introduced to help interpret the effects of the parameters on the shape of the regression function. Second, limiting (and beyond the limit) versions of the model are developed to clarify and extend the range of possible behaviors covered by the model. Finally, several issues related to estimation are discussed, including multimodality, singularity and reparameterization.

C0618: **Local-global extrinsic regression on manifolds**

Presenter: **Luca Maestrini**, The Australian National University, Australia
Co-authors: Janice Scealy, Francis Hui, Andrew Wood

Local-global extrinsic regression models are developed on manifolds that are similar in spirit to semiparametric regression models on Euclidean spaces. The local features are assumed to result from a general unknown function defined on the non-Euclidean space, which can be estimated using a smoothing method. The global components are modelled through a parametric regression where a link function maps linear combinations of regression coefficients and covariates onto the non-Euclidean space. It is shown that for non-Euclidean spaces with sufficiently rich isometry groups, such as spheres, it is possible to write the non-parametric and parametric components in the regression function as multiplicative factors.

This multiplicative structure can be exploited to efficiently fit the models via backfitting algorithms. Inference for the model parameters of interest can be carried out by exploiting unbiased estimating equations.

C0771: **Empirical likelihood on manifolds**
Presenter: **Andrew Wood**, Australian National University, Australia
Co-authors: Karthik Bharath, Huiling Le

Empirical likelihood (EL) is a type of nonparametric likelihood that is useful, e.g. for constructing nonparametric confidence regions in one-sample problems for intrinsic or extrinsic Frechet means and in k-sample testing problems for Frechet means, especially when one wishes to avoid the assumption of a common dispersion structure across populations. An important property of EL is that it obeys Wilk's theorem. EL has previously been developed for data on particular manifolds, including the unit sphere, with applications to directional statistics; real projective space, with applications to axial data; and complex projective space, with applications to (Kendall) shape data for objects represented by labelled landmarks in two dimensions. However, in previous work on EL, there has been no attempt to develop a general, unified approach to EL for general manifolds. Manifolds are treated with positive curvature and negative curvature separately, using extrinsic geometry in the former case and intrinsic geometry in the latter for reasons that will be explained. A unified approach to EL will be developed in each setting. EL is considered for data on Stiefel and Grassmann manifolds, for example. It turns out that EL is straightforward to implement for data from these and other manifolds. Moreover, Wilk's theorem holds in general manifold settings, and bootstrap calibration is available and, under mild conditions, has desirable higher-order properties, as in the Euclidean case.

C1074: **A rolled Gaussian process model for curves on manifolds**
Presenter: **Alfred Kume**, University of Kent, United Kingdom
Co-authors: Simon Preston, Karthik Bharath, Pablo Lopez-Custodio

Curves on manifolds arise as data in various applications, but there are few available statistical models suited to them. The main obstacle is the nonlinearity of manifolds, which makes it difficult to specify useful models. Hence, one strategy is to flatten the manifold and then exploit the flattened space for modelling, but how the flattening is done is crucial because it induces distortions. "Unrolling" and "unwrapping" are harnessed to flatten the manifold in a particular optimal way that minimizes distortion local to the data, opening the path to a class of random effect models that are easy to work with.

CO377 Room K2.41 EVIDENCE INTEGRATION AND TRIANGULATION FOR GLOBAL HEALTH RESEARCH Chair: Samuel Manda

C1232: **Estimating spatial distribution of HIV prevalence in South Africa from multiple survey data sources**
Presenter: **Bedilu Alamirie Ejigu**, Addis Ababa University, Ethiopia
Co-authors: Samuel Manda

Analyses of multiple HIV data sources can enable accurate prediction of HIV burden. This is critical to support policy decision-makers in making more effective public health interventions to contain HIV. Using calibration techniques and spatial statistics methods, HIV data is combined from three different health surveys to estimate improved spatial patterns in the HIV burden in South Africa. HIV prevalence is found to be disproportionately varied between districts in South Africa. Synthesizing multiple data sources is encouraged to provide a more accurate representation of the health burden.

C1558: **A Bayesian hierarchical hidden Markov model for infectious diseases time series**
Presenter: **Geoffrey Singini**, University of Malawi, Malawi
Co-authors: Samuel Manda

In infectious disease forecasting, hidden Markov statistical time series models are used to understand the distribution of the observed disease data conditional on the hidden states and the transitions between states. For example, in human immunodeficiency virus (HIV) disease, hidden states (viral latency or activation) could be associated with observed states (unsuppressed or suppressed viral load). Using hidden Markov models could help draw a complete picture of HIV development. However, in many applications, the observations of the disease progression within a subject may be correlated. Thus, estimating the hidden Markov model parameters between and within subjects could improve the model's capabilities. Moreover, the observed disease data could have been collected from different sources, necessitating a joint model of multiple data streams. A Bayesian hidden Markov model that incorporates subject-level and data source heterogeneity is developed. The proposed methodology is demonstrated with extensive simulation studies and an application to HIV time series data from multiple data sources in Malawi.

C1589: **Assessing methods for detecting outliers in meta-analysis**
Presenter: **Memory Makuta**, University of Malawi, Malawi
Co-authors: Samuel Manda

Outlier data sets in meta-analyses can have a substantial negative impact on the validity of empirical findings and the strength of the conclusions. Detecting outliers and the ways to handle them are crucial in ensuring the reliability of meta-analytic findings. The random effects variance shift model is discussed in comparison to the more common ways of outlier detection in meta-analyses. These methods will be analytically compared for their ability to identify outliers and offer insights into their strengths and limitations. The theoretical evaluation of methods is supplemented by extensive simulation studies. Potential outlier data sets are then identified in a meta-analysis of the prevalence of unhealthy food consumption among children aged 6 to 23 months using several Demographic and Health Survey datasets from 2010 to 2022 across sub-Saharan African countries

C1403: **Statistics integration of health survey data for estimating disease spatial patterns**
Presenter: **Samuel Manda**, University of Pretoria, South Africa

In a two-phasing health survey sampling, disease data may only be observed in a subsample (nested data). The reduced sample could only be used for robust estimates of the disease spatial patterns at high levels of administrative aggregation. However, disease spatial patterns are increasingly needed at local levels for public health decision-making. Statistical methods are considered for imputation of the disease data for the remaining sample or independent samples that collect auxiliary information (non-nested structure). Combining information from different sources to obtain an improved official health estimate and association could be desired by health policymakers to reduce public health costs. A spatial analysis of adult HIV in sub-Saharan Africa exemplifies the methodology.

CO241 Room S0.03 TOPICS IN MULTIVARIATE AND HIGH-DIMENSIONAL DATA Chair: Ritwik Sadhu

C0356: **Bayesian high-dimensional linear regression with sparse projection-posterior**
Presenter: **Samhita Pal**, North Carolina State University, United States
Co-authors: Subhashis Ghosal

A novel Bayesian approach is considered for estimation, uncertainty quantification, and variable selection for a high-dimensional linear regression model under sparsity. The number of predictors can be nearly exponentially large relative to the sample size. A conjugate normal prior is put, initially disregarding sparsity, but for making an inference, instead of the original multivariate normal posterior, the posterior distribution is used, induced by a map transforming the vector of regression coefficients to a sparse vector obtained by minimizing the sum of squares of deviations plus a suitably scaled 1-penalty on the vector. The resulting sparse projection-posterior distribution shows that contracts around the true value of the parameter at the optimal rate adapted to the sparsity of the vector. The true sparsity structure gets a large sparse projection-posterior probability. An appropriately recentered credible ball is further shown to have the correct asymptotic frequentist coverage. Finally, how the computational burden can be distributed to many machines is described, each dealing with only a small fraction of the whole dataset. A comprehensive simulation study is conducted under a variety of settings, and the proposed method is found to perform well for finite sample sizes. The method is implemented in an R package named sparseProj, and it is applied to the ADNI data, where the ADAS score is predicted based on selected gene-expression data.

C0380: **Regularized estimation and inference of sparse spectral precision matrices**
Presenter: **Navonil Deb**, Cornell University, United States
Co-authors: Amy Kuceyeski, Sumanta Basu

Estimation of the spectral precision matrix, an object of central interest in frequency-domain time series, is a key step in calculating partial coherency and graphical model selection of stationary time series. When the dimension of a time series is large, traditional estimators of spectral precision tend to be severely ill-conditioned, and one needs to resort to suitable regularization strategies involving optimization over complex variables. Existing regularization approaches either separately penalize real and imaginary parts of a complex number or use off-the-shelf optimization routines for complex variables that do not explicitly leverage the underlying sparsity structure of the problem. A complex graphical Lasso (CGLASSO) is proposed as an L1-penalized estimator of a spectral precision matrix based on local Whittle likelihood maximization. Fast pathwise coordinate descent algorithms are developed to implement CGLASSO on large dimensional time series data sets. The algorithm relies on a ring isomorphism between complex and real matrices that maps a number of optimization problems over complex variables to similar optimization problems over real variables. In addition, a framework is proposed for the inference of CGLASSO across different frequencies in high dimensional regimes. Error bounds are calculated for a de-biased CGLASSO estimators and demonstrate asymptotic normality supported with empirical performance.

C0400: **OT rank tests for heterogeneous non-parametric two-sample testing**
Presenter: **Ritwik Sadhu**, Cornell University, United States
Co-authors: Nilanjan Chakraborty, Trambak Banerjee

In many modern inference tasks, the data-generation process contains inherent source distribution heterogeneity, often resulting in multi-modality of the composite dataset on which statistical procedures must be performed. Traditional parametric methods tailored towards uni-modal data are inefficient when applied to such data, while non-parametric methods (including multivariate non-parametric methods) still hold some chance of success, at least for traditional inference questions such as testing equality of two distributions or independence of paired observations. However, data containing distributional heterogeneity opens up the chance to ask some more sophisticated inference questions, such as the testing for the presence of an entirely new component distributions in a subset of the data. For this testing problem, henceforth called the remodeling problem, a new test statistic is proposed, constructed using optimal transport (OT)-based multivariate ranks, which allows for asymptotically consistent testing with a correct asymptotic level under the assumption of well-separated component distributions in the null. To the best of knowledge, this is the first statistic in the literature for this problem with the aforesaid guarantees.

C1003: **On a fast and consistent test for equality of means for high-dimensional data**
Presenter: **Sayan Das**, Washington University in St. Louis, United States
Co-authors: Debraj Das, Subhajit Dutta

A two-sample test is proposed for high-dimensional means using a logistic regression framework. The crux of the method relies on identifying significant feature variables via LASSO (using the logit link) that can capture the population mean. The proposed test is based on the regression coefficients of these selected features, effectively reducing dimensionality. Theoretically, it is shown that the proposed test is consistent against a wide range of alternatives. In the ultra-high-dimensional regime, the test leverages the sparse structures of the means, making it computationally more efficient compared to other methods in the existing literature. Additionally, the method is extended to test for equality of means across multiple groups. Finite sample studies corroborate the theoretical findings and reveal the superiority of the tests compared to several existing tests.

CO155 Room S0.11 STATISTICAL METHODS FOR BRAIN IMAGING DATA Chair: Simon Vandekar

C0395: **Improving statistical power of multi-modal associations via de-variation**
Presenter: **Ruyi Pan**, University of Toronto, Canada

Understanding the interplay between different modalities of brain MRI data is crucial for unravelling the complexities of brain structure and function. Existing statistical association tests for two random vectors are often limited in fully capturing dependencies between modalities, particularly by overlooking correlation structures within each modality, leading to the potential loss of statistical power. A novel approach, termed de-variation, is proposed to address this limitation. De-variation is considered a simple yet effective preprocessing method that leverages a penalized low-rank factor model to capture within-modality dependencies. Theoretical analyses and simulation studies show (i) its powerful performance when within-modality correlations impact signal-to-noise ratios and (ii) its robustness when these are absent. De-variation is then applied to brain imaging-driven phenotypes (IDPs) derived from functional, structural, and diffusion MRI from the UK Biobank to show its promising performance.

C0698: **Whole brain connectivity estimation by GPU-enhanced Gaussian process**
Presenter: **Hakmook Kang**, Vanderbilt University, United States
Co-authors: Minjee Kim, Yuting Mei, Ilwoo Lyu, Alisa Zoltowski, Chris Fonnesbeck, Carissa Cascio

The previous Bayesian spatiotemporal model approach has been expanded to significantly reduce the computation burden by employing GPU (graphics processing unit) computing and the Gaussian process to model the intra-voxel spatial correlation in each ROI (region of interest). A Bayesian double-fusion technique is used for enhancing the estimation of whole brain resting state functional connectivity (FC) based on functional magnetic resonance imaging (fMRI) data between brain regions by using structural connectivity (SC) based on diffusion tensor imaging (DTI) data. Concurrently acquired two imaging data are simultaneously used for FC estimation, which allows precise investigation of the relationship between FC and SC or alterations in white matter microstructural integrity. This enhanced modeling approach is applied to investigate the nuances of functional network connectivity in individuals with autism, potentially uncovering significant insights into the neural underpinnings of the disorder.

C0938: **Nonparametric methods for analysis of brain cortical gradients**
Presenter: **Andrew Chen**, University of Pennsylvania, United States

Recent methodological advances describe the topological organization of the brain cortex as a continuous map called brain cortical gradients. These gradients are consistent with seminal research on brain functional organization, well-studied neurodevelopmental trajectories, and key multimodal brain metrics. However, statistical methods for the analysis of brain cortical gradients are limited and current approaches either ignore population variability or key properties of gradient data. Brain cortical gradients and methods for deriving gradients are first introduced. Then, the unique properties of this novel data type and the limitations of existing approaches are discussed. Finally, nonparametric hypothesis testing methods appropriate for gradient data are proposed. Application to the Philadelphia Neurodevelopmental Cohort reveals that the proposed methods can capture neurodevelopmental changes in gradients and differences between demographic groups. Potential extensions and statistical frameworks are explored for further methodological developments.

C1000: **A Bayesian latent factor model for curve alignment and covariate-dependent smoothing**
Presenter: **Aaron Scheffler**, University of California, San Francisco, United States

Disease progression can be tracked via a cascade of changes in biomarkers and clinical measurements over the disease time course. For example, in progressive neurodegenerative diseases (ND), such as Alzheimer's Disease, changes in biomarkers (neuroanatomical images, cerebrospinal fluid) may precede clinical measurements (cognitive batteries) by months or years. Viewing repeated measurements of biomarkers and clinical measurements as a multivariate time series composed of continuous and discrete values, successful modeling of disease progression balances capturing stereotypic patterns in disease progression across subjects with subject-level variability in timing, acceleration, and shape of disease progression trajectories. A Bayesian latent factor model is proposed for curve alignment and covariate-dependent smoothing of exponential family outcomes across the disease time course, allowing for the characterization of typical disease progression as well as heterogeneity in the timing, speed, ordering, and shape of disease progression at the population-level and at the subject-level via random effects structure that partitions phase and amplitude variance. The framework will accommodate continuous and count outcomes, allowing for the incorporation of measurements ranging from neuroimaging features to sensitive sub-scales of cognitive batteries. A working example is provided from patients experiencing progressive ND.

CO333 Room S0.12 METHODS AND MODELS FOR ENVIRONMENTAL AND ECOLOGICAL DATA II Chair: Domenico Vitale

C0341: **Modelling spatiotemporal point processes for environmental applications**

Presenter: **Nicoletta D Angelo**, Universita degli Studi di Palermo, Italy
Co-authors: Giada Adelfio

Patterns of points in three-dimensional space, as well as those occurring in both space and time, have growing attention in data description methods. Three-dimensional geometry is particularly important in geology and astronomy, while examples of spatio-temporal point patterns are available in epidemiology (e.g., for disease surveillance), seismology, and anatomy (e.g., video microscopy). Modelling three-dimensional point processes can serve as a convenient way to analyze such complex phenomena across various fields. The aim is to provide an overview of this topic, exploring the fundamental concepts of point processes, including practical applications. Specialized models for analyzing complex spatiotemporal point patterns are presented, specifically designed for the analysis of patterns that occur in a subset of the Euclidean space or on some specific linear network, such as roads of a city. The main aim of fitting such models is, indeed, to better understand the distribution and interaction of events in a three-dimensional space.

C0430: **Analyzing summer wildfire patterns in Italian municipalities using satellite data**

Presenter: **Crescenza Calculli**, University of Bari, Italy, Italy
Co-authors: Alessio Pollice

The current availability of remote sensing technologies and access to large amounts of high-resolution satellite imagery have significantly enhanced wildfire monitoring and facilitated the refinement of localized firefighting strategies. Utilizing the fire (thermal anomalies) indicator acquired from NASA/MODIS instruments, a comprehensive spatial analysis is proposed for wildfire patterns across the Italian municipalities for the entire summer season of 2023. Modeling and predicting the occurrence of wildfires represent a complex challenge due to the broad range of spatial and temporal scales involved in wildfire processes, contributing to their nonlinear nature. Therefore, one of the most promising advancements in wildfire assessment is integrating spatial data into models that address the change of support problem, allowing for the combination and aggregation of point- and area-level analyses. The proposed investigation of zero-inflated wildfire counts for the municipalities allows for capturing local variability, highlighting the specific impacts of environmental and socio-economic drivers on the distribution of wildfires. Finally, the difficulties involved in integrating and consolidating data are recognized for local and regional analyses. Acknowledging these challenges can increase the likelihood that future monitoring endeavors will be adapted or planned to better incorporate data from multiple sources.

C0427: **Heteroskedastic hidden dynamic geostatistical models for environmental data**

Presenter: **Jacopo Rodeschini**, University of Bergamo, Italy
Co-authors: Alessandro Fusta Moro, Andrea Moricoli, Alessandro Fasso

A common framework for studying spatiotemporal processes is the state space model (SSM) and the related Kalman filter technique. In this context, a well-established multivariate spatiotemporal model is the hidden dynamic geostatistical model (HDGM), which is an SSM suitable for complex environmental processes. The observation variability is modelled by the measurement equation, which is essentially given by a regression component, a stochastic latent process, and an error term. The spatiotemporal correlation is modelled by the latent equation through a Markovian process, with the innovation term being a zero-mean Gaussian process with a spatial covariance function. In environmental studies, addressing heteroskedasticity is crucial for accurate inference. In spatiotemporal models, heteroskedasticity can relate to time, space, data heterogeneity, or a combination thereof. Data heterogeneity often occurs in data fusion problems, where data come from various sensors or processes. Two heteroskedastic extensions of the HDGM are compared, featuring time-varying error variance. The first method treats error variance as a nuisance parameter, employing a flexible, unstructured, time-varying error variance. The second method models error variance as a linear combination of time-based basis functions. Both methods estimate model parameters using the expectation-maximization algorithm.

C0736: **Demystifying spatial confounding**

Presenter: **Emiko Dupont**, University of Bath, United Kingdom
Co-authors: Isa Marques, Thomas Kneib

Environmental and ecological data are often collected across geographical regions. Regression models for such spatially indexed data use spatial random effects to approximate unmeasured spatial variation in the response variable. However, as spatial random effects are usually not independent of the covariates in the model, this can lead to significant bias in the covariate effect estimates, making them unreliable. This fundamental problem, known as spatial confounding, has received considerable interest in recent years, not least because the established methods for dealing with it were proven to be problematic. However, spatial regression models are typically complex, and research into the topic has sometimes led to puzzling and seemingly contradictory results. A broad theoretical framework is developed that brings mathematical clarity to the mechanisms of spatial confounding, providing explicit analytical expressions for the resulting bias. It is seen that the problem is directly linked to spatial smoothing, and it is exactly identified how the size and occurrence of bias relate to the features of the model and the underlying confounding scenario. Using the results, subtle and counter-intuitive behaviors are explained, and a general approach is proposed for dealing with spatial confounding bias in practice, applicable for any spatial model specification.

CO289 Room S0.13 PERSONALIZED MEDICINE AND REINFORCEMENT LEARNING (VIRTUAL) Chair: Ruoqing Zhu

C1254: **Fiducial approaches to censored survival data**

Presenter: **Yifan Cui**, Zhejiang University, China
Co-authors: Jan Hannig

Novel nonparametric and semiparametric fiducial approaches to censored survival data are introduced. Gibbs samplers are proposed, and Bernstein-von Mises theorems are established. The estimators are also demonstrated by extensive simulations and real data applications.

C1255: **Dual active learning for reinforcement learning from human feedback**

Presenter: **Wei Sun**, Purdue University, United States

Aligning large language models (LLMs) with human preferences is critical to recent advances in generative artificial intelligence. Reinforcement learning from human feedback (RLHF) is widely applied to achieve this objective. A key step in RLHF is to learn the reward function from human feedback. However, human feedback is costly and time-consuming, so an effective strategy to collect human feedback within a sample budget is essential. Additionally, different teachers have different levels of rationality in various types of contexts, making it critical to query the most informative teachers for their preferences. A dual active reward learning policy is introduced for the simultaneous selection of contexts and teachers motivated by the idea of D-optimal design.

C1309: **Estimation of average treatment effect for survival outcomes with continuous treatment in observational studies**

Presenter: **Qi Zheng**, University of Louisville, United States

Co-authors: Triparna Poddar, Maiying Kong

In healthcare research, where extensive observational datasets such as claims data and electronic health records are abundant, researchers often aim to explore both the effects of treatments and the mechanisms by which these effects occur. While recent literature on causal effects in survival analyses typically concentrates on binary or multiple treatment scenarios, studies involving continuous treatment settings remain comparatively underexplored. Prompted by the need to assess the impact of blood lead levels on mortality among older adults in the United States, this project investigates the estimation of the average treatment effect (ATE) of continuous treatment on time-to-event outcomes. Estimating the ATE directly is proposed using an accelerated failure time-based marginal structural model (AFT-MSM). To tackle multiple confounding factors and censoring issues, the inverse probability of treatment weighting (IPTW) method is utilized, complemented by censoring weights. This approach has been rigorously validated through theoretical examinations and comprehensive simulation studies, affirming its validity and effectiveness. Additionally, the analysis suggests that the current regulatory level for blood lead is safe regarding mortality risk.

C1608: **Consistent order determination of Markov decision process**

Presenter: **Chuyun Ye**, Beijing Normal University, China

Co-authors: Lixing Zhu, Ruoqing Zhu

The Markov assumption in Markov decision processes (MDPs) is fundamental in reinforcement learning, influencing both theoretical research and practical applications. Existing methods that rely on the Bellman equation benefit tremendously from this assumption for policy evaluation and inference. Testing the Markov assumption or selecting the appropriate order is important for further analysis. Existing tests primarily utilize sequential hypothesis testing methodology, increasing the tested order if the previously-tested one is rejected. However, this methodology cumulates type-I and type-II errors in sequential testing procedures that cause inconsistent order estimation, even with large sample sizes. To tackle this challenge, a procedure is developed that consistently distinguishes the true order from others. A novel estimator is first proposed that equivalently represents any order Markov assumption. Based on this estimator, a signal function and an associated signal statistic are thus constructed to achieve estimation consistency. Additionally, the curve pattern of the signal statistic facilitates easy visualization, assisting the order determination process in practice. Numerical studies validate the efficacy of the approach.

CO094 **Room Safra Lec. Theatre** FUNCTIONAL AND DISTRIBUTIONAL DATA ANALYSIS **Chair: Sonja Greven**

C0351: **Functional quantile principal component analysis**

Presenter: **Jeff Goldsmith**, Columbia University, United States

Co-authors: Alvaro Mendez-Civieta, Ying Wei, Keith Diaz

Functional quantile principal component analysis (FQPCA) is introduced, a dimensionality reduction technique that extends the concept of functional principal components analysis (FPCA) to the examination of participant-specific quantile curves. The approach borrows strength across participants to estimate patterns in quantiles, and participant-level data is used to estimate loadings on those patterns. As a result, FQPCA is able to capture shifts in the scale and distribution of data that affect participant-level quantile curves and is also a robust methodology suitable for dealing with outliers, heteroscedastic data or skewed data. The need for such methodology is exemplified by physical activity data collected using wearable devices. Participants often differ in the timing and intensity of physical activity behaviors, and capturing information beyond the participant-level expected value curves produced by FPCA is necessary for robust quantification of diurnal patterns of activity. The methods are illustrated using accelerometer data from the National Health and Nutrition Examination Survey (NHANES) and produce participant-level 10%, 50%, and 90% quantile curves over 24 hours of activity. The proposed methodology is supported by simulation results and is available as an R package.

C0385: **A Bayes space point of view on climate warming**

Presenter: **Christine Thomas-Agnan**, CNRS, France

The impact of temperature warming is often analyzed with temperature temporal summaries, which can lead to a loss of valuable information. To address this issue, it is possible to take temperature into account as a functional parameter (function of time). An alternative and complementary point of view is put forward, which considers temperature distributions over time periods. Using the Bayes space formalism, the yearly distributions of maximum (or minimum) daily temperatures are analyzed across Vietnam's provinces over a 30-year period (1987-2016). The daily maximum temperatures are preprocessed using maximum penalized likelihood, resulting in density samples expressed on a B-spline basis. First, the presence of outlying densities is investigated both spatially and temporally with the ICS method (invariant component selection) adapted for density objects. Regional effects of the temperature density relative changes of these distributions between the initial period 1987-1989 and the final period 2014-2016 are examined using a Bayes-space version of functional analysis of variance. Lastly, the impact of climate warming is assessed on rice yield production using a scalar on density regression model.

C1149: **Can AI learn distributional regression**

Presenter: **Brian Caffo**, Johns Hopkins University, United States

Co-authors: Bonnie Smith

The challenges in having artificial intelligence in the form of deep learning is considered in learning invariances associated with distributional regression. Distributional regression is a particular challenge to learn in an automated fashion since the assumption of exchangeability of the covariate is factorial in the number of invariances. The cost of attempting to learn invariances versus an alternative strategy of assuming potential invariances is considered. Applications to biomedical data is used to illustrate results.

C0674: **Conformal uncertainty quantification using kernel depth measures in separable Hilbert spaces**

Presenter: **Marcos Matabuena**, Harvard University, Spain

Co-authors: Pavlo Mozharovskyi, Oscar Hernan Madrid Padilla, Jukka-Pekka Onnela, Rahul Ghosal

Depth measures have gained popularity in the statistical literature for defining level sets in complex data structures like multivariate data, functional data, and graphs. Despite their versatility, integrating depth measures into regression modeling for establishing prediction regions remains underexplored. To address this gap, a novel method is proposed utilizing a model-free uncertainty quantification algorithm based on conditional depth measures and conditional kernel mean embeddings. This enables the creation of tailored prediction and tolerance regions in regression models handling complex statistical responses and predictors in separable Hilbert spaces. The focus is exclusively on examples where the response is a functional data object. To enhance practicality, a conformal prediction algorithm is introduced, providing non-asymptotic guarantees in the derived prediction region. Additionally, both conditional and unconditional consistency results and fast convergence rates are established in some special homoscedastic cases. The model finite sample performance is evaluated in extensive simulation studies with different function objects as probability distributions and functional data. Finally, the approach is applied in a digital health application related to physical activity, aiming to offer personalized recommendations in the U.S. population-based on individuals' characteristics.

CO362 **Room BH (S) 1.01 Lec. Theatre 1** **ANALYSIS OF NON-STATIONARY TIME SERIES** **Chair: Yunyi Zhang**

C0483: **High-dimensional generalized penalized least squares**
Presenter: **Ilias Chronopoulos**, University of Essex, United Kingdom
Co-authors: Aikaterini Chrysikou, George Kapetanios

Inference is developed in high dimensional linear models with serially correlated errors. The Lasso estimator is examined under the assumption of a-mixing in the covariates and error processes. While the Lasso estimator performs poorly under such circumstances, it is estimated via GLS Lasso the parameters of interest and the asymptotic properties of the Lasso are extended under more general conditions. The theoretical results indicate that the non-asymptotic bounds for stationary dependent processes are sharper, while the rate of Lasso under general conditions appears slower as T,p. Further, debiasing methods are employed to perform inference uniformly on the parameters of interest. Monte Carlo results support the proposed estimator, as it has significant efficiency gains over traditional methods.

C0484: **Heterogeneous grouping structures in panel data**
Presenter: **Aikaterini Chrysikou**, Kings College, University of London, United Kingdom
Co-authors: George Kapetanios

The existence of heterogeneity is examined within a group, in panels with latent grouping structure. The assumption of within-group homogeneity is prevalent in this literature, implying that the formation of groups alleviates cross-sectional heterogeneity, regardless of the prior knowledge of groups. While the latter hypothesis makes inference powerful, it can often be restrictive. Models with richer heterogeneity that can be found both in the cross-section and within a group are allowed without imposing the simple assumption that all groups must be heterogeneous. The further contribution is to the method proposed in a prior study by showing that the model parameters can be consistently estimated and the groups, while unknown, can be identifiable in the presence of different types of heterogeneity. Within the same framework, the validity of assuming both cross-sectional and within-group homogeneity is considered using testing procedures. Simulations demonstrate the good finite-sample performance of the approach in both classification and estimation, while empirical applications across several datasets provide evidence of multiple clusters, as well as reject the hypothesis of within-group homogeneity.

C0507: **A nonparametric test for correlation between nonstationary time series: Addressing challenges with limited replicates**
Presenter: **Alex Yuan**, University of Washington, United States
Co-authors: Wenying Shou

In disciplines from ecology to neuroscience, researchers analyze correlations between pairs of nonstationary time series to infer interactions or shared influences among variables. This often involves testing whether an observed correlation is stronger than expected under the null hypothesis that time series are independent. With only one experimental replicate, testing for dependence between nonstationary time series is exceedingly challenging and generally requires strong assumptions. Conversely, with many replicates, a nonparametric trial-swapping permutation test can be used where within-replicate correlations are compared to between-replicate correlations. Although this test is largely assumption-free, its minimum achievable p-value is $1/n!$ (where n is the number of replicates), making significance thresholds like 0.05 unattainable when $n \leq 3$. A variant of this approach that can report lower p-values of $2/n^n$ or $1/n^n$ when there is strong evidence of dependence is described. This is useful for biomedical studies, where n is often $3 \sim 5$, limiting the significance obtained by permutation alone. The test prevents the false positive rate from exceeding the significance level and only requires that replicates are independent and identically distributed. The test is demonstrated by confirming the observation that groups of zebrafish swim faster when directionally aligned, using a public dataset with three biological replicates.

C0591: **Bootstrap-assisted inference for weakly stationary time series**
Presenter: **Yunyi Zhang**, The Chinese University of Hong Kong, Shenzhen, China
Co-authors: Efstathios Paparoditis, Dimitris Politis

The literature often adopts two types of stationarity assumptions in the analysis of time series, i.e., the weak stationarity, suggesting that the mean and the autocovariance function of a time series are time-invariant, and strict stationarity, indicating that the marginal distributions of the time series are time-invariant. While the strict stationarity assumption is vital from a theoretical aspect, it is hard to verify in practice. On the other hand, weak stationarity is relatively feasible, as it relies only on the second-order structures of the time series. Concerning this, statisticians may want to avoid relying on strict stationarity assumptions during statistical inference. The focus is on the analysis of quadratic forms within a weakly stationary time series. Specifically, it establishes the Gaussian approximation for quadratic forms of a short-range dependent weakly stationary time series. Building upon this result, the asymptotic distributions of the sample autocovariance, the sample autocorrelations, and the sample autoregressive coefficients are derived. Furthermore, it adopts the dependent wild bootstrap method to facilitate statistical inference. Numerical results verify the consistency of the proposed theories and methods. Strict stationarity is hard to ensure and verify for a real-life dataset. Therefore, the results should be able to assist statisticians in analyzing real-life time series.

CO238 **Room BH (SE) 1.01** **ADVANCES IN BAYESIAN VARIABLE SELECTION AND COMPUTING** **Chair: Vivekananda Roy**

C1210: **Generalized Markov chain importance sampling methods**
Presenter: **Quan Zhou**, Texas A&M University, United States

First, it is shown that for any multiple-try Metropolis algorithm, one can always accept the proposal and evaluate the important weight that is needed to correct for the bias without extra computational cost. This results in a general, convenient, and rejection-free Markov chain Monte Carlo (MCMC) scheme that extends beyond the conventional Metropolis-Hastings framework. Second, by further leveraging the importance sampling perspective on Metropolis-Hastings algorithms, an alternative importance sampling-based MCMC sampler is proposed on discrete spaces, along with a general theory on its complexity. Numerical examples suggest that the proposed algorithms are consistently more efficient than the original Metropolis-Hastings versions.

C1219: **Zero-order parallel sampling**
Presenter: **Giacomo Zanella**, Bocconi University, Italy
Co-authors: Francesco Pozza

Finding effective ways to exploit parallel computing in order to speed up MCMC convergence is an important problem in Bayesian computation and related disciplines. The zero-order (aka derivative-free) version of the problem is considered, where it is assumed that (a) the gradient of the target distribution is unavailable (either for theoretical, practical or computational reasons) and (b) the (expensive) target distribution can be evaluated in parallel at K different locations, and those evaluations can be used to speed up MCMC convergence. Two main contributions are provided in this respect. First, any method falling within a quite general multiple-proposal framework is shown to speed up convergence by log(K) factors in high dimensions. The fundamental limitation of such a framework, which includes multiple-try MCMC as well as many other previously proposed methods, lies in restricting possible moves to support the K evaluation points. The results are stated in terms of upper bounds on the spectral gap of the resulting scheme. Second, two ways are discussed (one based on stochastic gradient estimators and the other based on factorized proposals), which make better use of parallel computing and achieve polynomial speed-ups in K. Some of the methods share similarities, but also notable differences, with classical zero-order optimization methods.

C1221: **Preconditioning in Markov chain Monte Carlo**
Presenter: **Samuel Livingstone**, University College London, United Kingdom

The purpose is to discuss the quantification of the effectiveness of linear preconditioning in MCMC. Preconditioning is an attempt to modify a target distribution so that it is more amenable to sampling. Linear preconditioning is the most common choice and refers to the act of premultiplying the state vector by a constant matrix. Recent results on mixing times in MCMC are leveraged to show scenarios in which commonly used preconditioners will and will not improve sampling. Further discussion on designing subquadratic linear preconditioners that can perform well in the presence of high correlation may be made.

C1324: **Adaptive neighborhood methods for Bayesian variable selection and structure learning**
Presenter: **Jim Griffin**, University College London, United Kingdom

Some Bayesian methods lead to posterior distributions defined on large discrete spaces. Bayesian variable selection for regression models with p variables leads to a posterior distribution defined on a space with 2^p possible models. Bayesian structure learning in Gaussian graphical models leads to a posterior distribution on all possible DAGs. It is well understood that sampling from these posterior distributions is very challenging. The purpose is to review recent work on the use of adaptive random neighborhood with informed proposal (ARNI) samplers. These combine adaptive proposals with informed proposals. Informed proposals allow the methods to converge more quickly and mix better than vanilla MCMC methods and the adaptive proposal allows scaling to large p settings. The use of these methods is illustrated in generalized linear models and Gaussian graphical models.

CO338 Room BH (SE) 1.02 BAYESIAN METHODS FOR DATA WITH LATENT STRUCTURE	**Chair: Deborah Kunkel**

C0266: **A variational empirical Bayes approach to multivariate multiple regression, with applications to polygenic prediction**
Presenter: **Fabio Morgante**, Clemson University, United States
Co-authors: Peter Carbonetto, Gao Wang, Yuxin Zou, Abhishek Sarkar, Matthew Stephens

Multivariate (i.e., multi-outcome) multiple regression has been an important tool in different fields of applied statistics. One of these fields is quantitative genetics, where the aim is to accurately predict complex trait phenotypes from genotypes. Multivariate multiple regression can be used to predict multiple correlated phenotypes jointly from genotypes, leveraging the shared genetic effects across such phenotypes and improving accuracy over univariate analyses. However, effects can be shared across phenotypes in a variety of ways, so computationally efficient statistical methods are needed that can accurately and flexibly capture patterns of effect sharing. New Bayesian multivariate multiple regression methods are described as using flexible priors learned from the data, which are able to model many different patterns of effect sharing and specificity across outcomes. The methods are evaluated in their ability to predict multiple phenotypes from genotypes using simulations with different patterns of effect sharing across phenotypes as well as real data applications. The results show that these new methods can provide more accurate predictions than existing univariate and multivariate methods while also being computationally efficient.

C0511: **Understanding uncertainty in Bayesian clustering**
Presenter: **Sara Wade**, University of Edinburgh, United Kingdom
Co-authors: Cecilia Balocchi

The Bayesian approach to clustering is often appreciated for its ability to provide uncertainty in the partition structure. However, summarizing the posterior distribution over the clustering structure can be challenging. A prior study proposed to summarize the posterior samples using a single optimal clustering estimate, which minimizes the expected posterior variation of information (VI). In instances where the posterior distribution is multimodal, it can be beneficial to summarize the posterior samples using multiple clustering estimates, each corresponding to a different part of the space of partitions that receives substantial posterior mass. The aim is to propose finding such clustering estimates by approximating the posterior distribution in a VI-based Wasserstein distance sense. An interesting byproduct is that this problem can be seen as using the k-mediods algorithm to divide the posterior samples into different groups, each represented by one of the clustering estimates. Using both synthetic and real datasets, it is shown that the proposal helps to improve the understanding of uncertainty, particularly when the data clusters are not well separated, or when the employed model is misspecified.

C0958: **Incorporating heterogeneous types of uncertainty in small area estimates from multiple demographic data sources**
Presenter: **Emily Peterson**, Emory University, United States
Co-authors: Lance Waller

Hierarchical models for small-area estimation have a rich history within the statistical toolbox for analyzing national census data and demographic projections. With the advent of national-level surveys with regional components (e.g., the US American Community Survey, Census data, Demographic and Health Surveys) and data science-based estimates (e.g., WorldPop, Global Burden of Disease, Meta, and Google), there is an opportunity to incorporate multiple heterogeneous data sources to improve the accuracy of local small area estimates. While a hierarchical modeling framework often provides an approach for linking multiple types and layers of population data, there are important contrasts in data collection, data availability, and processing methodologies across data sources, such that each set of population counts may be subject to different sources and magnitudes of error. Firstly, a brief outline of types of US-based small area population estimates and associated errors is provided. Secondly, approaches to robustly fuse information are explored from multiple data sources to improve the accuracy of small area estimates and obtain associated uncertainties in order to provide an inferential framework for such estimates.

C1087: **Bayesian generalized linear models for correlated data with fewer latent variables**
Presenter: **Maryclare Griffin**, University of Massachusetts Amherst, United States

Many challenges arise when simulating Bayesian generalized linear model posterior distributions in practice, especially when the observed data is assumed to be dependent. The focus is on two challenges that stem from the introduction of one or more auxiliary latent variables for each observation. First, several popular methods for simulating from Bayesian generalized linear model posterior distributions rely on the introduction of an auxiliary random variable for each observation. These methods can scale poorly when the number of observations is large because they require additional posterior draws and repeated expensive matrix calculations. Second, many of the most useful approaches for introducing dependence in the observed data do so by introducing a latent random variable with a dense but computationally convenient prior covariance matrix. However, the computational conveniences offered by the prior covariance matrix may be absent (or appear to be absent) from the posterior. Methods are introduced to address these challenges that take advantage of simple reparameterizations of the problem, advances in posterior mode computation, and modern sampling.

CO048 Room BH (SE) 1.05 MACHINE LEARNING FOR FINANCIAL DATA	**Chair: Andrii Babii**

C0710: **Tensor PCA for factor models**
Presenter: **Andrii Babii**, University of North Carolina, United States
Co-authors: Eric Ghysels, Junsu Pan

Modern empirical analysis often relies on high-dimensional panel datasets with non-negligible cross-sectional and time-series correlations. Factor models are natural for capturing such dependencies. A tensor factor model describes the multidimensional panel as a sum of a low-rank component and idiosyncratic noise, generalizing traditional factor models for two-dimensional panels. Several algorithms are considered to estimate the factors and factor loadings for tensor factor models. The asymptotic distribution theory is provided, and a test for the number of factors is proposed in a tensor factor model. The asymptotic results are supported by the Monte Carlo experiments, and the new tools are applied to sorted portfolios.

C0704: **Estimation and inference for CP tensor factor models**
Presenter: **Bin Chen**, University of Rochester, United States
Co-authors: Yuefeng Han, Qiyang Yu

High-dimensional tensor-valued data have recently gained attention from researchers in economics and finance. The estimation and inference of high-dimensional tensor factor models are considered, where each dimension of the tensor diverges. The focus is on a factor model that admits CP-type tensor decomposition, which allows for non-orthogonal loading vectors. Based on the contemporary covariance matrix, an iterative simultaneous projection estimation method is proposed. The estimator is robust to weak dependence among factors and weak correlation across different dimensions in the idiosyncratic shocks. An inferential theory is established, demonstrating both consistency and asymptotic normality under relaxed assumptions. Within a unified framework, two eigenvalue ratio-based estimators are considered for the number of factors in a tensor factor model, and their consistency is justified. Through a simulation study and two empirical applications featuring sorted portfolios and international trade flows, the advantages of the proposed estimator are illustrated over existing methodologies in the literature.

C0974: **Data and model uncertainty in the cross-section of equity returns**
Presenter: **Jiantao Huang**, University of Hong Kong, Hong Kong
Co-authors: Serhiy Kozak

A two-layer hierarchical Bayesian framework is developed to study the effects of data and model uncertainty on the pricing of the cross-section of equity returns. In the first layer, which handles data uncertainty, a Bayesian Tensor Model is proposed to generate a probability distribution of missing, infrequently, or imprecisely observed characteristic data. It is then drawn from this distribution to estimate a Bayesian factor pricing model of stock returns, which can be seen as a probabilistic generalization of the IPCA model. Bayesian averaging across the space of factor pricing models provides regularization similar to a prior study. Further averaging across posterior draws of characteristics data provides additional robustness with respect to uncertainty in the characteristics data. Jointly, accounting for both sources of uncertainty more than doubles the Sharpe ratios of alpha portfolios.

C0992: **Stripping the discount curve: A robust machine learning approach**
Presenter: **Markus Pelger**, Stanford University, United States
Co-authors: Damir Filipovic, Ye Ye

A robust, flexible and easy-to-implement method is introduced for estimating the yield curve from Treasury securities. The non-parametric method learns the discount curve in a function space that we motivate by economic principles. An extensive empirical study on U.S. Treasury securities shows that the method strongly dominates all parametric and non-parametric benchmarks. It achieves substantially smaller out-of-sample yield and pricing errors while being robust to outliers and data selection choices. The superior performance is attributed to the optimal trade-off between flexibility and smoothness, which positions our method as the new standard for yield curve estimation.

CO329 Room BH (SE) 1.06 RECENT DEVELOPMENTS IN BAYESIAN METHODS Chair: Hee Cheol Chung

C0196: **Order-based structure learning without score equivalence**
Presenter: **Hyunwoong Chang**, Texas A&M University, United States

Structure learning of directed acyclic graph (DAG) models is the task of discovering the underlying DAG structure that represents the conditional independence relations among variables in a given observational data. An empirical Bayes formulation of the structure learning is proposed, where the prior specification assumes that all node variables have the same error variance, an assumption known to ensure the identifiability of the underlying causal DAG. To facilitate efficient posterior computation, the posterior probability of each ordering is approximated by that of a best DAG model, which naturally leads to an order-based Markov chain Monte Carlo (MCMC) algorithm. Strong selection consistency for the model in high-dimensional settings is proved under a condition that allows heterogeneous error variances, and the mixing behavior of the sampler is theoretically investigated. The method is demonstrated to outperform other state-of-the-art algorithms under various simulation settings, and a single-cell real-data study is provided to illustrate the practical advantages of the proposed method.

C0437: **Synergizing roughness penalization and basis selection in Bayesian spline regression**
Presenter: **Seonghyun Jeong**, Yonsei University, Korea, South
Co-authors: Sunwoo Lim

Bayesian P-splines and basis determination through Bayesian model selection are both commonly employed strategies for nonparametric regression using spline basis expansions within the Bayesian framework. Despite their widespread use, each method has particular limitations that may introduce potential estimation bias depending on the nature of the target function. To overcome the limitations associated with each method while capitalizing on their respective strengths, a new prior distribution is proposed that integrates the essentials of both approaches. The proposed prior distribution assesses the complexity of the spline model based on a penalty term formed by a convex combination of the penalties from both methods. The proposed method exhibits adaptability to the unknown level of smoothness while achieving the minimax-optimal posterior contraction rate up to a logarithmic factor. An efficient Markov chain Monte Carlo algorithm is provided to implement the proposed approach. The extensive simulation study reveals that the proposed method outperforms other competitors in terms of performance metrics or model complexity.

C0797: **Hierarchical Bayes nested error regression models with spatial random effects**
Presenter: **Hee Cheol Chung**, UNC Charlotte, United States

In many applications, the population means of geographically proximate small areas exhibit spatial variation. When auxiliary variables fail to adequately capture this spatial pattern, the residual variation is absorbed into the random effects, violating the assumption of independent and identically distributed random effects. This issue becomes more significant when predicting means for low or non-sampled small areas, as these predictions depend almost entirely on auxiliary variables. To address this, spatial nested error regression models are proposed that account for interdependencies between small areas by incorporating spatially correlated random effects.

C0847: **TopSpace: Bayesian spatial topic modeling for unsupervised discovery of spatial tissue structures in multiplex images**
Presenter: **Junsouk Choi**, University of Michigan, United States

The recent development of multiplex imaging technologies allows for measuring the expression of tens of protein markers at single-cell resolution while preserving spatial information of cells, enabling direct observation of cellular phenotypes, spatial distributions, and interactions in intact tissues. A key research question in analyzing such data is to identify higher-order patterns of tissue organization, which holds systematic implications for disease pathology and clinical outcomes. To address this, TopSpace, a novel Bayesian topic model, is proposed to identify the higher-order architecture of tissues and recover signatures of characteristic cellular microenvironments that are potential determinants of patient outcomes. The proposed approach infers the local distribution of cellular phenotypes to represent the cellular microenvironment and incorporates spatial information via Gaussian processes to ensure spatial coherence among neighboring microenvironments. By applying the proposed TopSpace to publicly available multiplexed imaging data, higher-order architectures are uncovered within lung cancer tissues and identified tertiary lymphoid structures, which are strongly associated with patient survival.

CO078 Room BH (S) 2.01 ESG AND FINANCIAL-ECONOMIC SUSTAINABILITY Chair: Caterina Morelli

C0667: **The greenness of European green bonds in an asset pricing setting**
Presenter: **Paola Galfrascoli**, University of Milano-Bicocca, Italy
Co-authors: Elisa Ossola, Gianna Monti

A systematic green risk factor is proposed to be used in an asset pricing model for bond instruments. This factor is evaluated as the difference in returns between a green and a brown portfolio, where greenness is measured by a synthetic indicator incorporating several dimensions contributing to the definition of greenness at the bond and the issuer level. In the first case, information is such as the presence of a green label attributed by the issuer itself and/or by a data provider based on the use of proceeds of the funds raised and certifications by external institutions. At the issuer level, the environmental rating provided by MSCI and industry classification are considered, identifying sectors that are climate-policy relevant. The focus is on the analysis on European corporate bonds for which such information is available in FactSet and Bloomberg databases.

C0897: **Analyzing the interplay between ESG dimensions and corporate performance using a multiplex network approach**
Presenter: **Saverio Storani**, Sapienza University of Rome, Italy

The primary objective is to analyze corporate sustainability policies' interaction and reciprocal influence, encompassing environmental, social, and governance (ESG) dimensions over time and their impact on corporate performance. Utilizing a high-quality historical dataset, it is examined how changes in one ESG dimension influence the other two and how these effects are reflected in overall corporate performance. A multiplex network approach is adopted, where each layer corresponds to one of the ESG dimensions. The networks within each layer are constructed based on the principle of similarity among firms, allowing us to analyze correlations in corporate performance within each stratum. The temporal interactions between ESG dimensions are also studied to observe potential causality and general spill-over effects, considering the entire multiplex network. This aims to identify the dynamics linking ESG scores to the performance of the companies under investigation. Furthermore, centrality analysis and classifying firms into subgroups based on their ESG scores explore how more effective sustainability policies can positively influence corporate performance. Finally, strategic recommendations are formulated to improve sustainability policies and optimize corporate performance.

C0990: **Innovators network and green firms: An analysis on the propensity of becoming a green innovator**
Presenter: **Claudia Sartirana**, University of Milano-Bicocca, Italy
Co-authors: Maria Luisa Mancusi, Bart Baesens

Environmental and green technologies research are fundamental pillars of sustainable growth. In particular, green innovations, more than other innovations, result from the combination of knowledge and competencies related to different domains. Hence, it is important to explore the knowledge flows and the network effects in green innovation dynamics. The contribution to the literature is via a focus on the effect of network centrality measures originating from a network of innovators on the propensity of firms to innovate in green technologies, using patent data and their citations as a measure of innovative activity and source of knowledge. Combining network analysis and statistical models, the results show that being closely connected to firms who previously invested in green technologies has a positive effect on becoming green. Moreover, being central in a network of innovators has a significant impact on the propensity to become green. More specifically, firms that can get easy access to other firms' knowledge have a higher probability of filing green patents, while firms holding highly cited patents, meaning they are already a key reference in their field, have a lower probability of starting patenting in green technologies.

C0890: **Pattern and dynamics of ESG scores of European firms: Spatiotemporal cluster analysis**
Presenter: **Caterina Morelli**, University of Milan Bicocca, Italy
Co-authors: Simone Boccaletti, Paolo Maranzano, Philipp Otto

The assessment of corporate sustainability performance is extremely relevant in facilitating the transition to a green and low-carbon intensity economy. However, companies located in different areas may be subject to different sustainability and environmental risks and policies. Henceforth, the main objective is to investigate the spatial and temporal pattern of the sustainability evaluations of European firms. A large dataset is leveraged, containing information about companies' sustainability performances, measured by MSCI ESG ratings and geographical coordinates of firms in Western Europe between 2013 and 2023. By means of a modified version of the hierarchical algorithm, a spatial clustering analysis is conducted, combining sustainability and spatial information, and a spatiotemporal clustering analysis, which combines the time dynamics of multiple sustainability features and spatial dissimilarities, to detect groups of firms with homogeneous sustainability performance. Cross-national and cross-industry clusters are built with remarkable differences in terms of sustainability scores. The findings help to capture the diversity of ESG ratings across Western Europe and may assist practitioners and policymakers in evaluating companies facing different sustainability-linked risks in different areas.

CO073 Room BH (S) 2.02 NEW FRONTIERS IN ARTIFICIAL INTELLIGENCE APPLICATIONS Chair: Emanuela Raffinetti

C0529: **Investigating the RGB approach for safe AI**
Presenter: **Emanuela Raffinetti**, University of Pavia, Italy
Co-authors: Paolo Giudici, Golnoosh Babaei

Artificial intelligence applications require the development of practical tools that can mitigate risks arising from their use. Specifically, in order to ensure the trustworthiness of artificial intelligence systems, the four main key principles of sustainability (robustness), accuracy, fairness, and explainability have to be achieved. Recently, a new methodology named "Rank graduation box" (RGB) was introduced as a unified approach which shares the same theoretical root and allows to overcome one of the main drawbacks of the existing methods, i.e. the high computational effort. The behavior of the RGB metrics is further explored by means of simulation experiments, which can be easily reproduced. The experimental results indicate that these metrics are easy to interpret and that they can be applied to any machine-learning model independently of the underlying data.

C0252: **The effect of foreign direct investments on CO2 emissions, evidence from Asia**
Presenter: **Alessandro Spelta**, University of Pavia, Italy

The nexus between foreign direct investments (FDI) and CO2 emissions in the Asian continent. Optimal transport theory and satellite data are leveraged to investigate the dynamics of the joint and conditional distributions of FDI and CO2 emissions of Asian countries and regions and to forecast the future patterns of such variables. The approach avoids assuming a specific parametric form for density distributions, providing a flexible representation of the relationships between FDI and CO2 emissions. Findings highlight the emergence of a positive relationship between FDI and CO2 emissions, while forecasts result in a generally increasing trend of the aggregate FDI and CO2. Different patterns of the predicted FDI and CO2 emissions are instead observed for a pool of selected countries.

C0219: **Forecasting composite indicators: The role of environmental variables**
Presenter: **Polinesi Gloria**, Universita Politecnica delle Marche, Italy
Co-authors: Maria Cristina Recchioni, Francesca Mariani, Mariateresa Ciommi

In recent years, there has been an increasing focus on evaluating well-being at the local level. Since 2013, the Italian Institute of Statistics (ISTAT) has annually published a dashboard of indicators to measure Equitable and Sustainable Well-Being (BES) for Italy, its macro-areas (NUTS-1), and regions (NUTS-2). More recently, ISTAT has provided BES indicators at the local level (NUTS-3) for the Italian provinces and metropolitan cities. The aim is to provide a more in-depth analysis of territorial inequalities and divergences across the Italian provinces. First, the main pillars of BES (economic, social, environmental, and others) are synthesized using the parameters (mode and concentration) of the beta distribution underlying multidimensional well-being. These parameters, used as proxies for territorial disparities, reveal a high degree of heterogeneity not only between Northern and Southern Italian provinces but also among neighboring ones. Along this line, to rank Italian provinces a composite indicator of well-being is constructed by using a machine learning approach. Additionally, a new measure is introduced to evaluate the importance of a single indicator in terms of how it affects Italian well-being. This measure reflects the complex and multidimensional nature of well-being, where environmental variables play a crucial role.

C0175: **Culture "profiling", AI and AML: Efficacy vs ethics**
Presenter: **Parvati Neelakantan**, Indian Institute of Technology Kanpur, India

Using extensive transaction and money laundering detection data at a globally important financial institution, the efficacy of including aspects of national culture is investigated in formulating anti-money laundering predictions. For corporate and individual accounts, Hofstede individualism scores of the country in which a customer is resident or from which a wire is sent/received are of first-order importance. When combined with account and transaction data, as well as even a proprietary institutional algorithm, individualism scores continue to determine the models' predictive performances. The efficacy finding of profiling in AML compliance underscores the need for stringent and enforced data protection safeguards, which can serve to ensure an individual's fundamental right to privacy.

CO216 Room BH (S) 2.05 SUBJECTIVE BELIEFS, SURVEYS, AND OPTIONS DATA Chair: Alberto Quaini

C0646: **Institutions' return expectations across assets and time**
Presenter: **Markus Ibert**, Copenhagen Business School, Denmark
Co-authors: Magnus Dahlquist

The purpose is to study the equity, Treasury bond, and corporate bond risk premium expectations of asset managers, investment consultants, wealth advisers, public pension funds, and professional forecasters. Subjective risk premia vary one-to-one with objective risk premia that are available in real-time and known to be countercyclical (i.e., high in recessions and low in expansions). Despite their significant countercyclical time-series variation, many subjective risk premia vary more in the cross-section than in the time series, indicating persistent heterogeneity. This heterogeneity in subjective risk premia is tied to heterogeneity in expectations about long-run valuation levels. Overall, the results support rational expectations of asset pricing models that generate countercyclical risk premia and heterogeneous expectations.

C0873: **Subjective risk and return**
Presenter: **Theis Jensen**, Yale University, United States

Traditional asset pricing models like the CAPM explain realized returns worse than newer asset pricing models like Fama-French-5, but why? The aim is to show that (i) traditional models are better at explaining the subjective risk of stock but that (ii) newer models are better at predicting return-enhancing mispricing and that (iii) these outcomes can be explained by a model in which all the CAPM assumptions hold, except that some investors have an optimism bias. The results suggest that the search for better models of realized returns has produced better models of mispricing but worse models of risk.

C0904: **Investor beliefs and trading frictions**
Presenter: **Sofonias Alemu Korsaye**, Johns Hopkins University, United States

A theoretical framework is developed to identify investors' subjective beliefs consistent with survey expectations and asset prices in markets with trading frictions. A metric is introduced to quantify the deviation of these beliefs from rational expectations (RE), interpretable as a bound on the difference between the maximum Sharpe ratios under investors' beliefs and RE. Empirically, it is shown that a significant share of the deviation from RE assessed, assuming frictionless markets, can be attributed to small trading costs. This deviation and the impact of trading costs differ across investor characteristics, with sophisticated investors' expectations more closely aligning with asset prices.

C1104: **The dynamics of subjective risk, risk premia and beliefs**
Presenter: **Alberto Quaini**, Erasmus University of Rotterdam, Netherlands
Co-authors: Sofonias Alemu Korsaye, Gustavo Freire

Asset pricing research traditionally assumes that investors form rational expectations, yet investors' market expectations extracted from survey data are at odds with those predicted by rational expectation models based on market data. This divergence signals a gap in existing asset pricing theories, which are still evolving to accommodate empirical evidence on how investors shape their beliefs. In response to this need, a novel framework is developed to extract information on investors' conditional subjective beliefs about future cash flows from asset prices and survey expectations. The extraction is adaptable, allowing for the integration of diverse economic assumptions ranging from entirely model-free constraints, like good-deal bounds, to fully model-based scenarios, such as those assuming log-utility investors. Empirically, the dynamics of investors' subjective risk, risk premia and beliefs are investigated.

CO190 Room BH (SE) 2.01 ADVANCES ON BAYESIAN BIOSTATISTICS AND BIOINFORMATICS (VIRTUAL) Chair: Marco Ferreira

C1005: **Genome-wide iterative fine-mapping for related individuals**
Presenter: **Marco Ferreira**, Virginia Tech, United States
Co-authors: Jacob Williams, Shuangshuang Xu

Current fine-mapping methods are often implemented on small SNP sets and require the maximum number of causal SNPs to be less than a prespecified value. Examining an entire genotype array with these methods is often not computationally possible and impractical as determining the maximum number of causal SNPs is non-trivial. However, by not examining the entire array, these methods can miss weaker causal signals and, thus, have diminished statistical power. To address this issue, the method Genome-wide Iterative fiNe-mApping (GINA) is presented. It is shown with an extensive simulation study that, when compared to currently used fine-mapping methods, the method GINA reduces the false discovery rate and increases the discovery rate of true causal variants. The application of GINA is illustrated with case studies on plant science and human health.

C1020: **Bayesian dynamic clustering factor models: Estimating subgroups and transitions**
Presenter: **Allison Tegge**, Virginia Tech, United States
Co-authors: Tsering Dolkar, Marco Ferreira, Hwasoo Shin

With the increased recognition of heterogeneity in clinical cohorts, there is an increased need to develop algorithms to identify subgroups. Motivated by this need, and with an application to longitudinal health data as a case study, novel Bayesian dynamic clustering factor models are proposed. It is assumed that participants are assigned to one of several clusters at each time point. Each cluster corresponds to a health state. In addition, the transitions are modeled among the different clusters using a hidden Markov model. A Markov chain Monte Carlo algorithm is developed to explore the posterior distribution of the cluster means and the cluster transition probabilities. Finally, model selection is developed to concomitantly choose the number of clusters and the number of factors.

C0950: **Genome-wide iterative fine-mapping for non-Gaussian data**
Presenter: **Shuangshuang Xu**, Virginia Tech, United States
Co-authors: Marco Ferreira, Jacob Williams, Allison Tegge

Fine-mapping seeks to identify causal variants in genomic regions of interest previously identified by genome-wide association studies (GWAS).

However, because fine-mapping is performed separately from GWAS, fine-mapping does not extract as much information as possible from the data. A novel genome-wide iterative fine-mapping method for non-Gaussian data (GINA-X) is presented. GINA-X efficiently extracts information from GWAS data by iterating two steps: a screening step and a model selection step. The screening step provides a list of candidate genetic variants and an estimate of the proportion of null genetic variants. After that, the model selection step searches the model space defined by the list of candidate genetic variants and uses the estimated proportion of null genetic variants to appropriately control for genome-wide multiplicity. A simulation study shows that, when compared to competing fine-mapping methods, GINA-X reduces the false discovery rate and increases recall of true causal genetic variants. The usefulness and flexibility of GINA-X are illustrated with two case studies on alcohol use disorder and breast cancer.

C0976: **Bayesian dynamic clustering factor models with regressors**
Presenter: **Tsering Dolkar**, Virginia Tech, United States
Co-authors: Marco Ferreira, Allison Tegge, Hwasoo Shin

A novel class of Bayesian dynamic clustering factor models with regressors are proposed. This new class of factor models is useful for the analysis of multivariate longitudinal data on a sample of subjects. It is assumed that at each time point, each subject belongs to one of many clusters, and the subject may move to another cluster at the next time point. Further, the probability of a subject moving from one cluster to the other clusters depends on regressors. These regressors may include changes in individual-level psychosocial factors from one time point to the next time point. A Markov chain Monte Carlo algorithm is developed to explore the posterior distribution of the unknown quantities. The usefulness of the novel framework is illustrated with an application to health data.

CO163 **Room BH (SE) 2.10** **NONPARAMETRIC, SEMIPARAMETRIC, AND FUNCTIONAL DATA ANALYSIS** **Chair: Jiaying Weng**

C0493: **Semi-supervised triply robust inductive transfer learning**
Presenter: **Mengyan Li**, Bentley University, United States
Co-authors: Tianxi Cai, Molei Liu

A semi-supervised triply robust inductive transfer learning (STRIFLE) approach is proposed, which integrates heterogeneous data from a label-rich source population and a label-scarce target population and utilizes a large amount of unlabeled data simultaneously to improve the learning accuracy in the target population. Specifically, a high dimensional covariate shift setting is considered, and two nuisance models are employed, a density ratio model and an imputation model, to combine transfer learning and surrogate-assisted semi-supervised learning strategies organically and achieve triple robustness. Different from double robustness, even if both nuisance models are misspecified when the shifted source population and the target population share enough similarities, the triply robust STRIFLE estimator can still partially utilize the source population, and it is guaranteed to be no worse than the target-only surrogate-assisted semi-supervised estimator with an additional error term from transferability detection. These desirable properties of the estimator are established theoretically and verified in finite samples via extensive simulation studies. The STRIFLE estimator is utilized to train a Type II diabetes polygenic risk prediction model for the African American target population by transferring knowledge from electronic health records linked to genomic data observed in a larger European source population.

C0752: **Dimension reduction through imbalanced learning**
Presenter: **Qin Wang**, The University of Alabama, United States

Sufficient dimension reduction (SDR) is a useful tool in high-dimensional data analysis. It aims to find informative embedding subspaces without losing regression information. Inverse regression-based methods, including SIR and SAVE, have been proposed and well-studied in the literature. A resampling-based approach through imbalanced learning is proposed to further enhance estimation accuracy and consistency and ease the challenge of selecting the number of slices using traditional SDR approaches. Numerical studies and a real data application will be presented to illustrate the efficacy of the proposed method.

C1017: **Functional sufficient dimension reduction through average Frechet derivatives**
Presenter: **Kuang-Yao Lee**, Temple University, United States
Co-authors: Lexin Li

Sufficient dimension reduction (SDR) embodies a family of methods that aim for the reduction of dimensionality without loss of information in a regression setting. A new method is proposed for nonparametric function-on-function SDR, where both the response and the predictor are a function. The notions of functional central mean subspace and functional central subspace are first developed, forming the population targets of our functional SDR. An average Frechet derivative estimator is then introduced, which extends the gradient of the regression function to the operator level and enables the development of estimators for the functional dimension reduction spaces. The resulting functional SDR estimators are unbiased and exhaustive, and more importantly, without imposing any distributional assumptions such as the linearity or the constant variance conditions that are commonly imposed by all existing functional SDR methods. The uniform convergence of the estimators for the functional dimension reduction spaces is established while allowing the number of Karhunen-Loeve expansions and the intrinsic dimension to diverge with the sample size. The efficacy of the proposed methods is demonstrated through both simulations and two real data examples.

C0931: **Sufficient dimension reduction for high-dimensional nonlinear vector autoregressive models**
Presenter: **Jiaying Weng**, Bentley University, United States
Co-authors: S Yaser Samadi

Vector autoregressive (VAR) models are fundamental tools for analyzing multivariate time series data across diverse domains. However, modeling high-dimensional time series data poses challenges due to the curse of dimensionality, especially when incorporating multiple time series and escalating model complexity. Sufficient dimension reduction (SDR) is a concept in statistics and machine learning aimed at finding a lower-dimensional subspace of the original feature space that preserves the relevant information for the target variable(s) or response variable(s). The SDR is explored in nonlinear vector autoregressive (NVAR) models, where the current state depends on multiple indices defined in past lags with an unknown relationship. A novel time series martingale difference divergence matrix (MDDM) approach is proposed, tailored for non-sparse estimation, albeit suitable only for low-dimensional scenarios. However, in the context of high-dimensional complexities where the dimensions grow rapidly by increasing the size of series and lag order, a sparse estimation procedure is designed within the proposed MDDM method, leveraging a regularized optimization framework equipped with the LASSO penalty. Theoretical foundations are rigorously presented for both non-sparse and sparse estimators. Through simulations and real data analyses, the efficacy of the methodology in handling high-dimensional time series data is demonstrated within NVAR frameworks.

CO258 **Room BH (SE) 2.12** **MEASUREMENT ERROR AND MISSING DATA IN STUDIES OF TIME AND SPACE** **Chair: Sarah Lotspeich**

C0492: **Connecting healthy food proximity and disease: Straight-line vs. map-based distances**
Presenter: **Sarah Lotspeich**, Wake Forest University, United States

Healthy foods are essential for a healthy life, but accessing healthy food can be more challenging for some people. This disparity in food access may lead to disparities in well-being, potentially with disproportionate rates of diseases in communities that face more challenges in accessing healthy food. Identifying low-access, high-risk communities for targeted interventions is a public health priority. Current methods to quantify food access rely on distance measures that are either computationally simple (like the shortest straight-line route) or accurate (like the shortest map-based driving route), but not both. A multiple imputation approach is proposed to combine these distance measures, harnessing the computational ease of one with the accuracy of the other. The approach incorporates straight-line distances for all neighborhoods and map-based distances for just a subset, offering comparable estimates to the "gold standard" model using map-based distances for all neighborhoods and improved efficiency over the "complete case" model using map-based distances for just the subset. Through a measurement error framework, straight-line distances are leveraged to impute for neighborhoods without map-based distances. Using simulations and data for North Carolina, U.S.A., the associations of diabetes and obesity are quantified with neighborhood-level proximity to healthy foods. Imputation also makes it possible to predict an area's full food access landscape from incomplete data.

C0580: **A Bayesian hierarchical model to account for temporal misalignment in American community survey explanatory variables**
Presenter: **Jihyeon Kwon**, Drexel University, United States
Co-authors: Staci Hepler, David Kline

The American Community Survey (ACS) is one of the most vital public sources for demographic and socioeconomic characteristics of communities in the United States and is administered by the U.S. Census Bureau every year. The ACS publishes 5-year estimates of community characteristics for all geographical areas and 1-year estimates for areas with a population of at least 65,000. Many epidemiological and public health studies use 5-year ACS estimates as explanatory variables in models. However, doing so ignores the uncertainty and averages over variability during the time period, which may lead to biased estimates of covariate effects of interest. A Bayesian hierarchical model is proposed that accounts for the uncertainty and disentangles the temporal misalignment in the ACS multi-year time-period estimates. It is shown via simulation that the proposed model more accurately recovers covariate effects compared to models that ignore the temporal misalignment. Lastly, the proposed model is implemented to quantify the relationship between yearly, county-level characteristics and the prevalence of frequent mental distress for counties in North Carolina from 2014 to 2018.

C0367: **Spatial filtering for unified calibration of air pollution data from multiple low-cost sensor networks**
Presenter: **Claire Heffernan**, Merck, United States

Low-cost air pollution sensors are increasingly being deployed worldwide, creating networks that provide information on local variability within a region, but these sensors have considerable measurement error. In some cities, including Baltimore, Maryland, there are multiple low-cost networks covering the same area, providing several sources of information about air quality. While there are many available methods to calibrate data from one low-cost network, separate calibration of each network leads to conflicting predictions of air quality. A unified Bayesian spatial filtering model is developed that combines data from multiple low-cost networks as well as any available reference devices in the region to provide unified predictions at any location within the region. The method allows for network-specific calibration equations as biases and noise levels of low-cost sensor networks depend on the type of sensor used since the measurement error varies across networks. Also, the method guards against potential preferential sampling of some of the networks, providing better predictions and narrower confidence intervals compared to calibrating each network individually. The method is fit to $\text{PM}_{2.5}$ data in Baltimore in June and July 2023. The approach can be used to calibrate low-cost air pollution sensor data from multiple sensor types in Baltimore going forward.

C0360: **Inferential challenges with spatial data in air pollution epidemiology**
Presenter: **Kayleigh Keller**, Colorado State University, United States

Many large-scale epidemiological studies investigate relationships between spatial and spatiotemporal exposures and adverse health outcomes. However, the spatiotemporal nature of these exposures can lead to inferential challenges, including measurement error and unmeasured spatial confounding. Spatiotemporal prediction of exposures induces errors that can be correlated across space and lead to bias in point estimates and standard errors of estimated health effects. Unmeasured factors that vary spatially and impact health can further cause confounding bias that is difficult to diagnose. The aim is to present methods for addressing both challenges in analyses of regional and national cohort studies of air pollution exposure and birth, cardiovascular, atopic, and cognitive health outcomes. The limitations of these correction approaches highlight important aspects of study design that can mitigate the effects of measurement error and unmeasured spatial confounding on inference.

CC466 Room S-1.04 STOCHASTIC PROCESSES Chair: Stefan Wrzaczek

C1090: **Adaptive Bayes estimator for stochastic differential equation with jumps under small noise asymptotics**
Presenter: **Shuntaro Suzuki**, Waseda University, Japan

Parameter estimation for stochastic differential equations driven by Wiener processes and compound Poisson processes is considered. Unknown parameters are assumed, corresponding to coefficients of the drift term, diffusion term, and jump term, as well as the Poisson intensity and the probability density function of the underlying jump. Estimators based on adaptive Bayesian estimation from discrete observations are proposed. The consistency and asymptotic normality of the estimators is demonstrated within the framework of small noise asymptotics.

C1246: **Quasi-Akaike information criterion of SEM for diffusion processes**
Presenter: **Shogo Kusano**, Osaka University, Japan
Co-authors: Masayuki Uchida

A model selection problem is considered for structural equation modeling (SEM) for diffusion processes. SEM is a statistical method that describes the relationships between latent variables. Statisticians often have some candidate models for SEM. In this case, selecting the optimal model among the competing models is necessary. Thus, many researchers have studied the information criteria of SEM for IID models. Recently, SEM for diffusion processes has been studied. The method enables the analysis of the relationships between latent processes based on high-frequency data. On the other hand, to the best knowledge, there are few studies on the information criteria for the SEM. Therefore, the quasi-Akaike information criterion (QAIC) of the SEM is proposed, and the asymptotic properties are investigated. The situation where a family of competing models includes some misspecified parametric models is also dealt with. It is proved that the probability of selecting the misspecified models by QAIC converges to zero. Furthermore, examples and simulation results are given to show the performance of the proposed information criterion.

C1405: **Multivariate additive subordination with applications in finance**
Presenter: **Giovanni Amici**, Politecnico di Torino, Italy
Co-authors: Laura Ballotta, Patrizia Semeraro

A tractable multivariate pure jump process is introduced in which the trading time is described by an additive subordinator. The multivariate process retains the additivity property and, therefore, is time inhomogeneous, i.e., its increments are independent but non-stationary. The theoretical framework provided for the process performs a sensitivity analysis with respect to the time inhomogeneity parameters and the design of a Monte Carlo scheme to simulate the trajectories of the process. The model is then employed in the context of option pricing in the FX market. The specific features of currency triangles are taken advantage of to extract the joint dynamics of FX log-rates. Extensive tests based on observed market data show that the model outperforms well-established pure jump benchmarks.

C1445: **Linear regression with Bouchaud's stochastic aging**
Presenter: **Andrej Srakar**, Institute for Economic Research Ljubljana, Slovenia

Aging is an out-of-equilibrium physical phenomenon gaining interest in physics and mathematics. Bouchaud has proposed the following toy model to study the phenomenon. Let $G = (\mathcal{V}, \mathcal{E})$ be a graph, and let $E = \{E_i\}_{i \in \mathcal{V}}$ be the collection of i.i.d. random variables indexed by vertices of this graph. The continuous-time Markov chain $X(t)$ is considered with state space $\mathcal{V}$. The transition rates w_{ij} are defined by $w_{ij} = \nu \exp\left(-\beta\left((1-a)E_i - aE_j\right)\right)$. Proving an aging result consists in finding a two-point function $F(t_w, t_w + t)$ such that a nontrivial limit $\lim_{t\to\infty(t/t_w)=\theta} F(t_w, t_w + t = F(\theta)$ exists. The aim is to introduce ageing in a linear regression model for cross-section and panel data. Appropriate regression specifications, estimation and inference procedures are proposed. Asymptotic results are provided based on the earlier literature on probability theory and simulation studies. In an application, the effects of childhood book reading and diseases on the health status of old-age individuals using retrospective panel models are studied. Including aging in regression models is novel and opens many unexplored possibilities. As Bouchaud's trap models are also underexplored in probability theory, this promises interesting avenues for research in econometrics and probability.

CC420 Room BH (S) 2.03 FINANCIAL ECONOMETRICS II Chair: Weining Wang

C1236: **Decomposing the term structure of credit spreads and predicting the macroeconomy in Japan**
Presenter: **Takeshi Kobayashi**, NUCB Business School, Japan

The purpose is to extract the common factors from the term structure of firm-based credit spreads of Japanese corporate bonds and examine the predictive content of the credit spread on the real economy. The dynamic Nelson-Siegel model is extended to allow for both common level, slope and curvature and firm-specific factors. The credit quality factors are also considered, which capture the difference between high-level and low-level credit spreads. The result shows that the estimated common and credit quality factors are important drivers of firm-based credit spreads and have substantial predictive power for future economic activity. The contribution to the literature is examining the relationship between firm-based credit spread curves and economic fluctuation and forecasting the business cycle.

C1258: **Analyzing the EPU effect on the risk of extreme events in the oil market: A MIDAS touch to dynamic POT models**
Presenter: **Katarzyna Bien-Barkowska**, Poznan University of Economics and Business, Poland
Co-authors: Agata Kliber, Rodrigo Herrera

The potential of the economic policy uncertainty (EPU) index is analysed to enhance the forecast of tail risk dynamics in the oil market. Specifically, a novel approach is proposed to harness the temporal clustering of such extreme events and their intensity in oil market returns by using a dynamic peaks-over-threshold (POT) model in discrete time that incorporates the EPU index within a MIDAS framework. The proposed model, referred to as the autoregressive conditional MIDAS-POT (AC-MIDAS-POT), combines a dynamic specification for inter-exceedance times and exceedance magnitudes, both of which are supplemented with MIDAS components to integrate macroeconomic information. Applied to Brent and WTI data, the findings reveal that increases in policy uncertainty significantly elevate the frequency and magnitude of extreme events, especially in the right tail of the distribution for Brent and, to a lesser extent, for WTI. The AC-MIDAS-POT model demonstrated superior risk forecasting performance compared to traditional GARCH-MIDAS models using Gaussian and Student t distributions, both in-sample and out-of-sample. This framework provides a robust tool for forecasting tail risk in oil markets and offers valuable insights for investors and policymakers to anticipate and mitigate the adverse effects of extreme price fluctuations.

C1674: **Closing the gap between state-space and score-driven models**
Presenter: **Xia Zou**, Vrije Universiteit Amsterdam, Netherlands
Co-authors: Andre Lucas, Yicong Lin

State-space and score-driven models are compared for option implied volatility surface dynamics. Point forecasts of both models behave similarly, but density forecasts of plain-vanilla score-driven models are substantially worse. This phenomenon is explained, and it shows how a simple adjustment of the measurement density of the score-driven model can put both models back on an equal footing. The score-driven models can subsequently be easily extended with non-Gaussian features to better fit the data without complicating parameter estimation. The findings are illustrated using S&P500 index options implied volatility surfaces.

C1398: **Recent extensions of short- and long memory volatility and duration models implemented with R**
Presenter: **Oliver Kojo Ayensu**, Paderborn University, Germany
Co-authors: Yuanhua Feng, Dominik Schulz

Recent advancements in short- and long-memory extensions of volatility (GARCH) and autoregressive conditional duration (ACD) models and their implementation in R are explored to enhance the understanding of dynamic behaviours in financial returns and non-negative time series. The study begins with a brief review of established GARCH and ACD models, then examines novel long memory GARCH variants, including fractionally integrated Log-GARCH (FILog-GARCH), modulus asymmetric FILog-GARCH (MAFILog-GARCH), and members of the "double power modulus EGARCH class", such as modulus modified EGARCH (MEGARCH) and modulus Log-GARCH (MLog-GARCH). Additionally, two new ACD models, namely the adjusted FIACD and long-memory Box-Cox ACD, are introduced. Details on their theoretical properties, estimation methods, and practical applications are investigated. A comprehensive review of relevant R packages is provided, offering practical guidance for implementing established and emerging models. This study serves as a valuable resource for researchers and practitioners seeking to apply advanced volatility modeling techniques using R.

CC447 Room BH (SE) 2.05 APPLIED ECONOMETRICS Chair: Nicola Loperfido

C1449: **Minimum wage and health status in Europe**
Presenter: **Sylwia Roszkowska**, Jagiellonian University, Poland
Co-authors: Aleksandra Majchrowska

The aim is to analyse the relationship between the minimum wage and the health status and mortality in European countries and regions. On the one hand, higher minimum wages can improve individuals' health by reducing financial stress and increasing consumption and investment in health. The significance of both the level of minimum wage and its changes and regional differences is analysed to explain the inter- and intraregional differences in health status and mortality across European countries. The analysis uses the classic methods of panel data analysis as well as those allowing for endogeneity, including the general method of moments estimator. A wide range of control variables are used, following the concepts of a past study and the rainbow model of health determinants proposed by another study, showing the importance of lifestyle and broadly defined socio-economic, cultural and environmental factors in health status. The results indicate that in countries with higher minimum-to-average wage ratios, the health status is higher, and the mortality is lower due to lower income inequalities. Moreover, the health status is higher, and the mortality rate is lower in countries with lower regional disparities in the minimum to average wage ratio.

C1307: **Treatment effect estimation in high-dimension: An inference-based approach**
Presenter: **Ulrich Aiounou**, Aix-Marseille School of Economics, France

Post-lasso and post-double lasso are becoming the most popular methods for estimating average treatment effects from linear regression models with many covariates. However, these methods can suffer from substantial omitted variable bias in finite samples. Autometrics, another variable selection method based on statistical inference, is considered and shown with simulation evidence that post-double autometrics performs well when the other methods fail and is illustrated in an application.

C0255: **Comparative analysis of the efficiency of health systems in OECD countries (2000-2020)**
Presenter: **Maria Rosa Nieto Delfin**, Investigaciones y Estudios Superiores, S.C, Mexico
Co-authors: Odra Angelica Saucedo Delgado, Juan Sebastian Valades-Garcia

The purpose is to examine the efficiency of health systems in OECD member countries, focusing on the Mexican health system from 2000 to 2020 and a forecast of its evolution until 2024. The objective is to contribute to the debate on more inclusive and efficient right-to-health policies and to inform the formulation of more effective initiatives in Mexico and other OECD countries. In a context where 36.6% of the population in Mexico lacked access to health services in 2020 and has the lowest investment in health relative to GDP in the OECD, it is understood how different health systems influence the optimization of resources and the quality of care. Dualistic, universal, and residual health systems are compared and forecasted using the dynamic data envelopment analysis (DEA) technique to assess how the distribution of resources affects the efficiency of health systems in OECD countries and to explore practices that could improve equity and efficiency in the provision of health services. The results indicate that the health system in Mexico is less efficient than that of other countries with dualistic health systems. Conversely, the empirical evidence indicates that the universal health system is the most efficient among OECD countries.

C1374: **Dynamic panel data models for the potential European crime drop**
Presenter: **Ilka van de Werve**, VU Amsterdam, Netherlands
Co-authors: Siem Jan Koopman

A panel data model is formulated with time-varying trends to empirically verify the possible existence of the European crime drop. The stochastically time-varying component represents the cross-national crime trend, and each country has its own weight on it. By representing the model in state space form, a likelihood-based approach is proposed using Kalman filtering to estimate the parameters and extract the unobserved time-varying component. In the second step, a cluster analysis of the country's fixed effects and weights is used to show (dis)similarities across the European countries. Several representations of the model are compared to the two-way fixed effects model and dynamic factor model. An empirical illustration of the presence of a potential European crime drop in comparison with the US crime drop shows the benefits of the proposed model and estimation methodology.

CC498 Room BH (SE) 2.09 FORECASTING II Chair: Antoine Djogbenou

C1546: **Option-implied physical distributions**
Presenter: **Richard McGee**, University College Dublin, Ireland
Co-authors: Valerio Poti, Thierry Post

The physical probability distribution of one-month equity index returns is forecasted using an initial density forecast, bid-ask prices of a cross-section of monthly index options, and systems of asset pricing conditions for incomplete options markets with frictions. The option pricing kernel is restricted to be a positive, monotonic, and/or convex function of index return, resembling the intertemporal marginal rate of substitution for standard utility functions. A physical density forecast is obtained by information projection of the initial forecast onto the set of distributions that are consistent with the prices and restrictions. The implied physical significantly improves upon the initial by using forward-looking information contained in the option prices. It also improves upon the implied risk-neutral, which confounds the physical density with the pricing kernel. The improvements in forecasting ability translate to an annual information ratio for growth optimal portfolios of well above 0.60 during both calm markets and volatile markets. The convexity condition appears crucial: relaxing it often leads to pathological kernel shapes, the convergence of the implied to the initial, and a sharp deterioration of the forecasting ability. By contrast, the monotonicity condition seems less relevant, and the forecasting ability actually benefits from allowing U-shaped kernels, especially during volatile markets.

C1661: **Nonlinear forecasting with many predictors using mixed data sampling kernel ridge regression models**
Presenter: **Farrukh Javed**, Lund University, Sweden
Co-authors: Kristofer Mansson, Deliang Dai, Peter Karlsson

Policy institutes such as central banks need accurate forecasts of key measures of economic activity to design stabilization policies that reduce the severity of economic fluctuations. Therefore, the recommendation is kernel ridge regression in a mixed data sampling framework. Kernel ridge regression can handle many predictors with a nonlinear relationship to the target variable. Consequently, it can potentially improve the currently used principal component-based methods when the economic data follow a nonlinear factor structure. In a Monte Carlo study, it is shown that the kernel ridge regression approach is superior in terms of mean square error and is more robust than principal component-based methods for different nonlinear data-generating processes. By using a dataset consisting of 24 economic indicators, Swedish gross domestic production is forecasted. The results confirm the superiority of the kernel ridge regression approach, especially during the economic crisis caused by the COVID-19 pandemic. Therefore, it is suggested that policy institutes consider the use of kernel-based approaches when forecasting key measures of economic activity.

C1378: **Geometric deep learning for realized covariance matrix forecasting**
Presenter: **Andrea Bucci**, University of Macerata, Italy
Co-authors: Michele Palma, Chao Zhang

Traditional methods employed in matrix volatility forecasting often overlook the inherent Riemannian manifold structure of symmetric positive definite matrices, treating them as elements of Euclidean space, which can lead to suboptimal performance and difficulties in parameter interpretation. Moreover, they often struggle to handle high-dimensional matrices. A novel approach for forecasting realized covariance matrices of asset returns is proposed using a Riemannian-geometry-aware deep learning framework. In this way, the geometric properties of the covariance matrices account for both possible non-linear dynamics and efficient handling of high-dimensionality. The efficacy of the approach is demonstrated using daily realized covariance matrices for the 20 most capitalized companies in the S&P 500 index, showing that the method outperforms traditional approaches in terms of forecasting accuracy.

C1613: **Forecasting dynamic correlation via a hybrid deep learning: Multivariate DCC GARCH model**
Presenter: **Yasemin Ulu**, Easten Michigan University, United States

The forecasting performance of the multivariate DCC -GARCH model is compared to that of a hybrid multivariate GARCH deep learning model for the stocks in the BIST30 index. Specifically, a hybrid model based on the recurrent deep neural network (RDNN) and DCC-GARCH models is used. The results are compared to that from a multivariate DCC-GARCH model. The results indicate that the hybrid model that utilizes both DCC-GARCH and deep learning combination performs better than the multivariate GARCH (DCC) model.

Monday 16.12.2024 16:50 - 18:30 Parallel Session Q – CFECMStatistics2024

CO147 Room S-2.25 DESIGN AND ANALYSIS OF STUDIES IN KIDNEY DISEASE (VIRTUAL) Chair: Jarcy Zee

C0650: **Target trial emulation to assess the effect of starting dialysis versus continuing medical management**
Presenter: **Maria Montez Rath**, Stanford University, United States
Co-authors: I-Chun Thomas, Vivek Charu, Michelle Odden , Carolyn Dacey, Shipra Arya , Enrica Fung, Ann OHare, Susan Wong , Manjula Kurella Tamura

For older adults who are not candidates for kidney transplantation, medical management is an alternative to lifelong dialysis, but little evidence is available to inform treatment decisions. As destination treatment, dialysis is intended to extend life and palliate symptoms. However, it also exposes patients to potential harms. In view of the potential risks and burdens of lifelong dialysis, medical management without dialysis is an alternative patient-centered treatment strategy. The national health care databases of the Department of Veterans Affairs are used to emulate a target trial of dialysis versus medical management in older adults with incident kidney failure who were not candidates for kidney transplantation. The restricted mean survival times are compared, and the results are contrasted to the expected number of days at home within three years of trial entry. Inverse probability weighting is used to estimate the analog of intention-to-treat and per-protocol analyses. Evidence of a modest survival benefit of starting dialysis is found compared to continuing medical management at the expense of fewer days at home, underscoring the importance of engaging patients in shared decision-making.

C0680: **Identifying new clinical trial surrogate endpoints in rare diseases: The PARASOL approach**
Presenter: **Abigail Smith**, Northwestern University, United States
Co-authors: Margaret Helmuth, Laura Mariani

Reasonably likely surrogate endpoints in rare diseases are challenging to develop due to a lack of data to provide regulatory confidence in the prediction of clinical benefit. In focal segmental glomerulosclerosis (FSGS), a rare glomerular disease affecting children and adults, endpoints are needed to support the development of new therapies. Proteinuria and GFR as clinical trial endpoints in focal segmental glomerulosclerosis (PARASOL) is an international collaborative effort to integrate observational, registry, and clinical trial data to define relationships between short-term changes in key disease activity and long-term clinical outcomes. Twenty-two datasets from around the world have been identified for data sharing. Challenges in integrating datasets include variation in the measurement of key biomarkers and the precision of documentation of a diagnosis of FSGS. This variation compounds underlying disease heterogeneity in age, underlying etiology, and rate of progression. Once combined, survival analysis approaches, including time-dependent and landmarked models, are applied to assess the ability of eGFR and UPCR to predict kidney failure in subgroups of patients with FSGS. In addition, clinical and data-driven approaches are used to define potential endpoints for trials, and sample size needs are assessed. Results from this project inform feasible clinical trial design and regulatory pathways for developing new therapies for FSGS.

C1114: **Feature selection and outcome prediction using kidney pathomic data**
Presenter: **Jarcy Zee**, University of Pennsylvania, United States
Co-authors: Qian Liu, Jeremy Rubin, Fan Fan, Laura Barisoni, Andrew Janowczyk

Kidney biopsy remains the gold standard for diagnosis and aids in prognostication for kidney diseases. Computational pathology leverages deep learning and automated image analysis technologies to quantitatively and comprehensively extract features from histological structures within digital biopsy images. Statistical analyses of the resulting pathomic data are challenged by their high-dimensional, hierarchical, and unbalanced nature. Kidney tubule pathomic data is used from the Nephrotic Syndrome Study Network (NEPTUNE) and Cure Glomerulonephropathy (CureGN) study to identify important morphologic features and predict clinical outcomes among patients with glomerular diseases. Two sets of analyses were conducted: first, tubule-level features were aggregated to the patient level using first- through fourth-order summary statistics, principal component analysis was used to group highly correlated features, minimum redundancy maximum relevance was used to rank feature groups, and a series of ridge regressions was used to select top features and predict survival outcomes; second, a novel CLUstering Structured lasSO (CLUSSO) scalar-on-matrix regression technique was developed and applied, which used cluster analysis to group similar tubules based on feature values and then used structured lasso to select important features and predict kidney function. These approaches allow for the selection of interpretable and clinically relevant pathogenic features to inform prognosis.

C0959: **Heterogeneous treatment effect estimation for longitudinal outcomes**
Presenter: **Vivek Charu**, Stanford University, United States
Co-authors: Tianyu Pan, Lu Tian

Developing trials around clinical endpoints for chronic kidney diseases, such as the event-time outcomes of end-stage kidney diseases, is often challenging due to the low likelihood of occurrence likeliness within ten years post-treatment. A slope-based endpoint, the rate of decline in eGFR in the first three years post-treatment, has been suggested as an effective surrogate for clinical endpoints. Existing research has focused on relaxing linear trend assumptions on eGFR and non-informative censoring assumptions or analyzing treatment effects in predetermined subgroups. Yet, none have explored data-driven sub-group identification and heterogeneous treatment effect estimation. The aim is to propose a method that introduces a Bayesian decision tree structure into a shared-parameter model, combining a survival model with a two-slope spline model that characterizes the rate of decline in eGFR. The proposed model simultaneously estimates the eGFR slope in the presence of informative censoring and provides digestible clinical decisions for sub-grouping governed by slope-based treatment effect heterogeneity. Simulation studies showcase that our proposed model effectively captures subtle heterogeneity in slope-based treatment effects. The model is also applied to the modification of diet in renal disease (MDRD) trial, providing Bayesian evidence that the patients with higher kidney failure risk score benefit more from the treatment.

CO388 Room S-1.04 NEW INVESTIGATORS IN COMPUTATIONAL STATISTICS (VIRTUAL) Chair: Rob Deardon

C0267: **Design of posterior analyses with sampling distribution segments**
Presenter: **Luke Hagar**, McGill University, Canada
Co-authors: Nathaniel Stevens

To design trustworthy Bayesian studies, criteria for operating characteristics of posterior analyses - such as power and the type I error rate - are often defined in clinical, industrial, and corporate settings. These operating characteristics are typically assessed by exploring entire sampling distributions of posterior probabilities via simulation. A scalable method is proposed to determine optimal sample sizes and decision criteria that maps posterior probabilities to low-dimensional conduits for the data. The method leverages this mapping and large-sample theory to explore sampling distributions of posterior probabilities in a nonuniform, targeted manner. This approach, based on exploring segments of sampling distributions, prompts consistent sample size recommendations with fewer simulation repetitions than standard methods. The posterior probabilities are repurposed, computed in that approach to efficiently investigate various sample sizes and decision criteria using contour plots.

C0299: **An integrated population model to estimate the population size of persons contending with homelessness**
Presenter: **Gracia Dong**, University of Toronto, Canada
Co-authors: Laura Cowen

Open population capture-recapture methods have been used with electronic healthcare records to estimate the population size of hidden vulnerable human populations, such as persons contending with homelessness. These methods rely on various assumptions about the data, which are often violated when data is not complete. Data privacy concerns can also limit access to computational resources, necessitating the need for more efficient model fitting or alternative modelling approaches. Integrated population models, which jointly analyse survey data from point-in-time counts and capture-recapture data from electronic healthcare records, can be used to provide more effective estimates. Electronic healthcare records spanning 2013-2023 are used from the Vancouver Island Health Authority, British Columbia, to identify adults in Greater Victoria contending with homelessness based on their self-reported housing status and the locations of the healthcare services used. Additionally, point-in-time counts (2018, 2020) are incorporated, which offer a snapshot of the homeless population on specific days. These counts include a breakdown by demographic data, proportions of individuals with underlying health concerns, and information on healthcare services accessed within the past year, including the proportion of individuals with emergency room visits.

C0506: **Computational efficiency and precision for replicated-count and batch-marked hidden population models**
Presenter: **Matthew Parker**, Simon Fraser University, Canada
Co-authors: Laura Cowen, Jiguo Cao, Lloyd Elliott

Two computational issues are addressed, common to open-population N-mixture models, hidden integer-valued autoregressive models, and some hidden Markov models. The first issue is computation time, which can be dramatically improved through the use of a fast Fourier transform. The second issue is the tractability of the model likelihood function for large numbers of hidden states, which can be solved by improving the numerical stability of calculations. As an illustrative example, the application of these methods is detailed in the open-population N-mixture models. Computational efficiency and precision are compared between these methods and standard methods employed by state-of-the-art ecological software. Faster computing times are shown (a 6 to 30 times speed improvement for population size upper bounds of 500 and 1000, respectively) over state-of-the-art ecological software for N-mixture models. The methods are applied to compute the size of a large elk population using an N-mixture model, and it is shown that while the methods converge, previous software cannot produce estimates due to numerical issues. These solutions can be applied to many ecological models to improve precision when logs of sums exist in the likelihood function and to improve computational efficiency when convolutions are present in the likelihood function.

C0954: **Multievent dynamic capture-recapture model: Estimating undetected COVID-19 cases in British Columbia, Canada**
Presenter: **Kehinde Olobatuyi**, University of Victoria, Canada
Co-authors: Patrick Brown, Laura Cowen

The accurate quantification of the impact of the COVID-19 pandemic on both public health and the economy is essential for informed policy-making. However, the true scope of the pandemic remains challenging to ascertain due to undetected cases, particularly when relying on reported cases, which rely heavily on test availability and strategies. To accurately quantify COVID-19 cases in British Columbia (BC), a Susceptible-Infectious-Recovered-multi-event capture-recapture (SIRMECR) model is developed to capture the dynamics of COVID-19. Specifically, the number of undetected COVID-19 cases is estimated in five health authority regions in BC, Canada, in 2020. Individual-level information available from the population data BC database is utilized to estimate the case detection probability, infection probability, survival probability, and recovery probability by incorporating testing volumes as covariates that improve the estimate of the parameters. A Markov chain Monte Carlo (MCMC) algorithm is developed to estimate MECMR model parameters. To address the computational challenges encountered, divide-and-conquer strategies are developed. The application provides an estimate of the total COVID-19 burden in the year 2020 and found the percentage of undetected varying from 77.4% to $84.0%.

CO400 **Room S-1.06** **GEOMETRIC DATA ANALYSIS FOR COMPLEX DATA STRUCTURES** **Chair: Carlos Soto**

C0325: **Analyzing spatial dependence in functional data and shapes of 2D curves**
Presenter: **Ye Jin Choi**, The Ohio State University, United States
Co-authors: Sebastian Kurtek, Karthik Bharath

The shapes of spatially dependent functional data or boundaries of two-dimensional (2D) objects, i.e., spatially dependent shapes of parameterized curves. Functional data is often composed of two confounded sources of variation: amplitude and phase. Amplitude captures shape differences among functions, while phase captures timing differences in these shape features. Similarly, boundaries of 2D objects represented as parameterized curves exhibit variation in terms of their shape, translation, scale, orientation and parameterization. The spatial dependence is studied among functions or curves by first decomposing given data into the different sources of variation. The proposed framework leverages a modified definition of the trace-variogram, which is commonly used to capture spatial dependence in functional data. Different types of trace-variograms are proposed to capture different components of variation in functional or shape data and use them to define a functional/shape mark-weighted K function by considering their locations in the spatial domain as random. This statistical summary then allows studying the spatial dependence in each source of variation separately. The efficacy of the proposed framework is demonstrated through extensive simulation studies and real data applications.

C0987: **A topological approach to analyzing access to resources with heterogeneous quality**
Presenter: **Sarah Tymochko**, University of California, Los Angeles, United States

Ideally, all public resources (e.g. parks, grocery stores, hospitals, etc.) should be distributed in a way that is fair and equitable to everyone. However, this is not always the case. Quantifying how much (or little) access individuals have to certain resources is a complex problem. Previous work has shown that tools from topological data analysis (TDA) can be useful in determining "holes" in the locations of resource locations based on geographical locations and travel times. Some resources may necessitate the incorporation of a notion of quality. As a case study, public parks are observed, which are heterogeneous in many ways. Having access to a park that is hundreds of acres with basketball courts, baseball diamonds, and an aquarium is inherently different than having access to a small patch of grass with an overgrown tennis court. An exploration of the access to public parks in Chicago is presented using persistent homology, a tool from TDA.

C0989: **Functional Gaussian differential privacy for private human faces**
Presenter: **Carlos Soto**, University of Massachusetts Amherst, United States

The problem of releasing a Gaussian Differentially Private (GDP) 3D human face is considered. The human face is a complex structure with many features and is inherently tied to one's identity. Protecting this data in a formally private way is important yet challenging, given the dimensionality of the problem. Approximate DP techniques for functional data is extended to the GDP framework. A novel representation, face radial curves, of a 3D face is further proposed as a set of functions, and then the proposed GDP functional data mechanism is utilized. To preserve the shape of the face while injecting noise, tools from shape analysis for the novel representation of the face are relied on. It is shown that the method preserves the shape of the average face and injects less noise than traditional methods for the same privacy budget. The mechanism consists of two primary components; the first is generally applicable to function value summaries (as are commonly found in nonparametric statistics or functional data analysis), while the second is general to disk-like surfaces and hence more applicable than just to human faces.

C1027: **Quantifying imaging heterogeneity via density functions with applications in brain and pancreatic cancer imaging**
Presenter: **Shariq Mohammed**, Boston University, United States

The quantification of heterogeneity in the tumor microenvironment is extremely important as visually differentiating disease types is challenging. A statistical framework is presented that quantifies spatial interactions in biomedical images to build prediction models for clinical phenotypes. A spatial regression model is first built to assess spatial interactions in regions of interest in the image. The heterogeneity in the spatial interactions in each image is then represented as a probability density function (serving as a signature quantifying spatial interactions). These density functions are analyzed using a Riemannian-geometric framework to include them as covariates in models that predict clinical outcomes of interest. The methodology is presented with applications to radiology imaging in brain cancer to predict isocitrate dehydrogenase mutation and to pathology imaging in pancreatic cancer to distinguish between different pancreatic disease subtypes.

CO021 Room S-1.27 RECENT ADVANCES OF MACHINE LEARNING IN INTERDISCIPLINARY PROBLEMS Chair: Tianxi Li

C1029: Selective inference after community detection on a single network

Presenter: **Daniel Kessler**, University of North Carolina at Chapel Hill, United States
Co-authors: Ethan Ancell, Daniela Witten

Networks arise in numerous applications in the social and biological sciences. Many network modeling tasks involve learning how to partition nodes into communities. While much work has focused on the statistical properties of community detection for edge-independent models, inference on the connectivity properties within and among communities has received relatively less attention. One particular challenge is that in many applications, only a single network is observed, and so both community detection and inference on just one network are conducted. Failing to account for this "double-dipping" can yield an invalid inference, but sample-splitting is nontrivial. By characterizing community detection as a form of model selection, recent developments are leveraged in selective inference in order to develop procedures for valid inference on the statistical properties of a single network after community detection. Because communities are learned from the data, there is the possibility of model misspecification, which is addressed using "sandwich estimators" of the variance. In general, the approach affords control of the so-called "selective Type I error rate" and is applicable to edge-independent networks with binary edges (using data fission) as well as many classes of weighted edges (using data thinning). Central limit theorems are established for the estimators, and the utility of the methods is demonstrated in numerical simulations.

C1566: Algorithms and incentives in machine learning

Presenter: **Haifeng Xu**, University of Chicago, United States

A generic question in statistics is to design approaches that take data as input and output estimation of certain parameters or prediction of some quantities. The standard paradigm often assumes these data are objectively generated from distributions without being affected by any human factors. However, this paradigm ceases to be true when the predictions or estimated parameters will, in turn, affect the data providers' welfare. In such situations, data providers have incentives to alter the data for their own benefit. Thus, the design of any statistical methods must account for potential data manipulations due to data providers' incentives. A general "incentive-aware" framework is introduced for designing prediction methods. This design paradigm is illustrated with two examples: (1) a very recent and timely application of eliciting authors' truthful private information for improving the peer review systems for today's massive scale machine learning conferences; (2) a very classic problem of PAC-learning classifiers but with strategic providers of data features. In both problems, the presence of incentives is illustrated to fundamentally change the problem's statistical efficiency and how algorithms can help to overcome some statistical barriers.

C1596: Isotonic mechanism for exponential family estimation in machine learning peer review

Presenter: **Yuling Yan**, University of Wisconsin-Madison, United States

In 2023, the International Conference on Machine Learning (ICML) required authors with multiple submissions to rank their submissions based on perceived quality. The aim is to employ these author-specified rankings to enhance peer review in machine learning conferences by extending the isotonic mechanism to exponential family distributions. This mechanism generates adjusted scores that closely align with the original scores while adhering to author-specified rankings. Despite its applicability to a broad spectrum of exponential family distributions, implementing this mechanism does not require knowledge of the specific distribution form. It is demonstrated that an author is incentivized to provide accurate rankings when her utility takes the form of a convex additive function of the adjusted review scores. For a certain subclass of exponential family distributions, it is proven that the author reports truthfully *only if* the question involves only pairwise comparisons between her submissions, thus indicating the optimality of ranking in truthful information elicitation. Moreover, it is shown that the adjusted scores dramatically improve estimation accuracy compared to the original scores and achieve nearly minimal optimality. A numerical analysis of the ICML 2023 ranking data is concluded with, showing substantial estimation gains in approximating a proxy ground-truth quality of the papers using the isotonic mechanism.

C1685: Flexible regularized estimating equations: Some new perspectives

Presenter: **Archer Yang**, McGill University, Canada

Some observations about the equivalences between regularized estimating equations, fixed-point problems and variational inequalities are made: (a) A regularized estimating equation is equivalent to a fixed-point problem, specified via the proximal operator of the corresponding penalty; (b) A regularized estimating equation is equivalent to a (generalized) variational inequality. Both equivalences can extend to any estimating equations with nonconvex penalty functions. To solve large-scale regularized estimating equations, it is worth pursuing computation by exploiting these connections. While fast computational algorithms are less developed for regularized estimating equations, there are many efficient solvers for fixed-point problems and variational inequalities. In this regard, we apply some efficient and scalable solvers which can deliver a hundred-fold speed improvement. These connections can lead to further research in both computational and theoretical aspects of the regularized estimating equations. We will also discuss applications of our approach in dynamic treatment regimes, instrumental variable regression and generalized estimation equations.

CO100 Room Auditorium MACHINE LEARNING IN ASSET PRICING Chair: Markus Pelger

C0558: Conditional latent factor models via model-based neural networks

Presenter: **Hao Ma**, Queen Mary University of London, United Kingdom

A hybrid methodology incorporating an econometric identification strategy is developed into artificial neural networks when studying conditional latent factor models. The time-varying betas are assumed to be unknown functions of numerous firm characteristics, and the statistical factors are population cross-sectional OLS estimators for given beta values. Hence, identifying betas and factors boils down to identifying only the function of betas, equivalent to solving a constrained optimization problem. For estimation, neural networks are constructed and customized to solve the constrained optimization problem, which gives a feasible non-parametric estimator for the function of betas. Empirically, the analysis is conducted on a large unbalanced panel of monthly data on US individual stocks with around 30,000 firms, 516 months, and 94 characteristics. It is found that 1) the hybrid method outperforms the benchmark econometric method and the neural networks method in terms of explaining out-of-sample return variation, 2) betas are highly non-linear in firm characteristics, 3) two conditional factors explain over 95% variation of the factor space, and 4) hybrid methods with literature-based characteristics (e.g., book-to-market ratio) outperform ones with COMPUSTAT raw features (e.g., book value and market value), emphasizing the value of academic knowledge from an angle of man vs. machine.

C0652: On the consumption wealth return

Presenter: **Paul Schneider**, USI Lugano and SFI, Switzerland
Co-authors: Paul Whelan, Marc Van Uffelen

The consumption wealth return is the fundamental state variable in long-run risk models. As the discrete-time continuous-state long-run-risk economy is highly incomplete, the consumption wealth return is not uniquely determined in equilibrium, however. Extant literature uses log-linearization, or log-polynomial frameworks, largely guided by tractability rather than economic principles. The consumption wealth return that minimizes the Hansen-Jagannathan bound nonparametrically is estimated. The resulting asset pricing statistics starkly contrast those of existing approaches, and the empirical results strongly underline the need for economic principles in the specification and estimation of economic models. In fact, some qualitative predictions that were thought to be model-implied in long-run risk models after using log-linearization turn out to vanish under the minimum-variance specification.

C1033: **Portfolio choice with unsystematic risk**
Presenter: **Paolo Zaffaroni**, Imperial College London, United Kingdom

A normative theory is developed for constructing mean-variance (MVE) portfolios, including the global-minimum-variance (GMV) portfolio, when unsystematic risk is compensated in a no-arbitrage setting. Two inefficient portfolios are identified: a portfolio that depends only on systematic risk factors and a portfolio that depends only on unsystematic risk, which, when combined, give mean-variance efficient portfolios. It is shown that, as the number of assets increases, the unsystematic risk portfolio" dominates the systematic risk portfolio" both in terms of the magnitude of its weights and in terms of its Sharpe ratio. A penalized estimator is then derived under a no-arbitrage constraint that allows for the econometric identification of the unsystematic risk portfolio, leading to a shrinkage-type estimator and demonstrating that it is equivalent to the robust portfolio" desired by investors averse to model misspecification. Finally, it is demonstrated that the theoretical insights lead to an economically and statistically significant improvement in out-of-sample performance.

C1217: **Firm characteristics and the cross-section of stock returns: A tale of two tails**
Presenter: **Daniele Bianchi**, Queen Mary University of London, United Kingdom
Co-authors: Pedro Moravis Venturi

The role of firm characteristics is explored to predict the cross-section of stock returns through the lens of a flexible Bayesian variable selection prior to being embedded in an otherwise conventional parametric portfolio choice. The main results show that model uncertainty is pervasive, and there is little evidence in favor of sparse models. Yet, there is a trade-off between sparsity and shrinkage when maximizing the portfolio's expected utility: while a heavy-tailed sparsity-inducing prior reduces uncertainty on which firm characteristics matter, it also produces strikingly less diversified portfolios with more extreme weights. As a result, when transaction costs are factored in, a dense model that allows for selecting many characteristics while shrinking their impact on the optimal portfolio choice is more adequate to capture the out-of-sample variation of stock returns.

CO037 Room K0.16 STATISTICAL ANALYSIS OF NETWORK AND OTHER COMPLEX DATA Chair: Yunpeng Zhao

C0201: **Nonparametric inference for balance in signed networks**
Presenter: **Weijing Tang**, Carnegie Mellon University, United States

In many real-world networks, relationships often go beyond simple dyadic presence or absence; they can be positive, like friendship, alliance, and mutualism, or negative, characterized by enmity, disputes, and competition. To understand the formation mechanism of such signed networks, the social balance theory sheds light on the dynamics of positive and negative connections. In particular, it characterizes the proverbs, "a friend of my friend is my friend" and "an enemy of my enemy is my friend". A nonparametric inference approach is proposed for assessing empirical evidence for the balance theory in real-world signed networks. The generating process of signed networks is first characterized with node exchangeability, and a nonparametric sparse graphon model is proposed. Under this model, confidence intervals are constructed for the population parameters associated with balance theory and establish their theoretical validity. The inference procedure offers higher-order accuracy and is more computationally efficient than bootstrap-based methods. By applying the method, strong real-world evidence for balance theory in signed networks across various domains is found, extending its applicability beyond social psychology.

C0295: **Higher-order accurate two-sample network inference and network hashing**
Presenter: **Yuan Zhang**, The Ohio State University, United States

Two-sample hypothesis testing for network comparison presents many significant challenges, including leveraging repeated network observations and known node registration but without requiring them to operate; relaxing strong structural assumptions; achieving finite-sample higher-order accuracy; handling different network sizes and sparsity levels; fast computation and memory parsimony; controlling false discovery rate (FDR) in multiple testing; and theoretical understandings, particularly regarding finite-sample accuracy and minimax optimality. A comprehensive toolbox is developed, featuring a novel main method and its variants, all accompanied by strong theoretical guarantees, to address these challenges. The method outperforms existing tools in speed and accuracy, and it is proven power-optimal. Algorithms are user-friendly and versatile in handling various data structures (single or repeated network observations; known or unknown node registration). An innovative framework has also been developed for offline hashing and fast querying, which is a very useful tool for large network databases. The effectiveness of the method is showcased through comprehensive simulations and applications to two real-world datasets, which revealed intriguing new structures.

C0410: **Identifying key influencers using an egocentric network-based randomized design**
Presenter: **Zhibing He**, Yale University, United States

Many public health interventions are implemented in settings where individuals are interconnected, and the intervention assigned to randomly selected individuals may also affect others within their network. Evaluating such interventions requires assessing both the effect of the intervention on those who receive it and the spillover effect on those connected to the treated individuals. With behavioral interventions, spillover effects can be heterogeneous in that certain individuals, due to their social connectedness and individual characteristics, are more likely to respond to the intervention and influence their peers' behaviors. Targeting these individuals can enhance the effectiveness of interventions in the population. An egocentric network-based randomized trial (ENRT) design is proposed, where a set of index participants is recruited from the population and randomly assigned to the treatment group while concurrently collecting outcome data on their untreated network members. In such a design, an estimator is developed to assess heterogeneous spillover effects and a testing method, the multiple comparison with best (MCB), to identify subgroups whose treatment exhibits the largest spillover effect on their network members.

C1028: **Enhancing bankruptcy prediction: A two-layered network approach using latent space models**
Presenter: **Tianhai Zu**, University of Texas at San Antonio, United States

A novel statistical approach is presented to corporate bankruptcy prediction by leveraging complex network analysis. A two-layered network structure is introduced that captures both supply chain relationships and investment-co-investment patterns among companies, providing a more comprehensive view of corporate interdependencies than traditional methods. To analyze this complex structure, a flexible multi-layered latent position model is developed that efficiently extracts key features from the network. The methodology employs advanced statistical techniques to estimate latent positions underlying this two-layered network, which are then utilized as predictors in a bankruptcy prediction model. Using the US public company data, it is demonstrated that incorporating these network-derived features significantly enhances the predictive power of bankruptcy models. The results reveal that these latent positions estimated from network structure capture crucial relational information that is highly relevant to a company's financial stability. This approach not only outperforms traditional prediction methods but also provides interpretable insights into the role of corporate interconnectedness in financial risk. The aim is to offer a robust statistical framework for integrating complex relational data into predictive modeling for bankruptcy risk assessment.

CO019 Room K0.18 METHODOLOGY AND PRACTICE FOR DATA ORIGINATING FROM RANDOMIZED TRIALS Chair: Andrew Spieker

C1231: Exploring the nature of individualized treatment effects using a large crossover trial

Presenter: **Bryan Blette**, Vanderbilt University Medical Center, United States

Methods for assessing individualized treatment effects (namely, conditional average treatment effects estimated for each individual in a study sample) have become increasingly popular for characterizing treatment effect heterogeneity in clinical trials. The interpretation of these effects is non-standard, and they are frequently conflated with more narrowly defined individual treatment effects. While individual treatment effects are never known, within-individual differences in a gold-standard crossover trial can provide a close analogue to individual effects. Individualized treatment effects are estimated using a suite of machine learning methods and data from the first stage of a relatively large crossover trial studying the effect of high- vs low-sodium diet on blood pressure. The second stage of the crossover trial is then revealed, and these estimates are compared to the within-individual differences observed in the trial. Through this illustration and under certain assumptions, subtle differences are highlighted between individualized and individual treatment effects, quantifying the extent to which the estimated individualized treatment effects would or would not accurately characterize heterogeneity of effect of the study treatment.

C1253: Experimenting with finite to infinite populations

Presenter: **Jonathan Chipman**, University of Utah, United States

Co-authors: Oleksandr Sverdlov, Diane Uschner

ANOVA is a common inferential strategy for randomized trials and assumes observations are drawn from an infinite population. However, trial participants are often considered a finite population based on the inclusion/exclusion criteria, location of the trial, and single-point timing of the trial. A finite population central limit theorem provides a degree of reassurance to assume normality when carrying out complete randomization with fixed equal allocation. Yet, in practice, many trials restrict randomization to reduce the risk of chronological bias by using a maximum tolerable imbalance procedure (MTI) or permuted block design (PBD). The impact on Type I error when using MTI or PBD for a finite population is not well studied. Through extensive simulations, common restrictions to randomization are observed to impact the Type I error convergence rate. When using ANOVA in a finite population, Type I error is more well controlled when implementing complete randomization than when using MTI or PBD. Randomization-based inference ensures an exact 5% inference under all settings and is reflective of the population from which patients are drawn.

C1297: A Bayesian approach to studying major adverse cardiovascular events: Leveraging information from clinical trials

Presenter: **Amber Hackstadt**, Vanderbilt University Medical Center, United States

Co-authors: Cara Lwin, Robert Greevy, Kathryn Snyder, Christianne Roumie

Multiple meta-analyses of trials estimated that the risk of major adverse cardiovascular event and heart failure hospitalization outcomes (MACE+HF) was lower for patients treated with sodiumglucose cotransporter 2 inhibitors (SGLT2i) versus dipeptidyl peptidase 4 inhibitors (DPP4i). However, the results were more varied in cohorts of patients without a history of cardiovascular disease (primary prevention). A Bayesian approach is applied to further investigate the association of SGLT2i with MACE+HF in a large primary prevention cohort of veterans. The Bayesian approach allows straightforward incorporation of prior information from other studies, including clinical trials. A Bayesian survival analysis model is used where the covariates are directly modeled in the hazard function, and a flexible M-spline function is used to estimate the baseline hazard. Information is incorporated from previous studies via an informative prior on the coefficient for the treatment effect in the hazard function. The Bayesian approach allows the estimation of the probability of a protective effect of SGLT2i for the MACE+HF outcome and its components. Different choices are examined for the priors. The Bayesian analysis suggested a protective effect for SGLT2i versus DPP4i for the MACE+HF outcome but not all the components of the MACE+HF outcome.

C1469: Causal mediation analysis of engagement for randomized trials involving mobile health interventions

Presenter: **Andrew Spieker**, Vanderbilt University Medical Center, United States

Clinical studies of mobile health interventions have received attention in recent years. A key challenge that emerges in these studies is understanding the role of engagement in driving the effects of these interventions (e.g., response rate to text message-delivered interventions). Some of the key nuances associated with both instrumental variable and mediation-based analyses of engagement in clinical trials of mobile health interventions are discussed. In particular, key topics will include (1) the blessings and curses of strong access monotonicity, (2) the applicability of the exclusion restriction, and (3) the applicability of pure and total direct/indirect effects. These matters are discussed in the context of recent studies of text message-delivered interventions such as REACH, FAMS, and VERB.

CO166 Room K0.19 STATISTICAL INNOVATIONS IN CLINICAL TRIAL DESIGN AND ANALYSIS Chair: Chenguang Wang

C0546: BF-BOIN-ET: A backfill Bayesian optimal interval design using efficacy and toxicity outcomes for dose optimization

Presenter: **Kentaro Takeda**, Astellas Pharma, United States

The primary purpose of a dose-finding trial for novel anticancer agents is to identify an optimal dose (OD), defined as the tolerable dose that has adequate efficacy in unpredictable dose-toxicity and dose-efficacy relationships. The FDA project Optimus reforms the paradigm of dose optimization and recommends that dose-finding trials compare multiple doses to generate these additional data at promising dose levels. The backfill is helpful in settings where the efficacy of a drug does not always increase with the dose level. More information is available at these doses by backfilling patients at lower doses while the trial continues to explore higher doses. A Bayesian optimal interval design is proposed using efficacy and toxicity outcomes that allow patients to be backfilled at lower doses during a dose-finding trial while prioritizing the dose-escalation cohort to explore a higher dose. A simulation study shows that the proposed design, the BF-BOIN-ET design, has advantages compared to the other designs in terms of the percentage of correct OD selection, reducing the sample size, and shortening the duration of the trial in various realistic settings.

C0549: On the mixed-model analysis of covariance in cluster-randomized trials

Presenter: **Bingkai Wang**, University of Michigan, United States

In the analyses of cluster-randomized trials, mixed-model analysis of covariance (ANCOVA) is a standard approach for covariate adjustment and handling within-cluster correlations. However, when the normality, linearity, or random-intercept assumption is violated, the validity and efficiency of the mixed-model ANCOVA estimators for estimating the average treatment effect remain unclear. Under the potential outcomes framework, the mixed-model ANCOVA estimators for the average treatment effect are proven consistent and asymptotically normal under arbitrary misspecification of its working model. If the probability of receiving treatment is 0.5 for each cluster, it is further shown that the model-based variance estimator under mixed-model ANCOVA1 (ANCOVA without treatment-covariate interactions) remains consistent, clarifying that the confidence interval given by standard software is asymptotically valid even under model misspecification. Beyond robustness, several insights on precision are discussed among classical methods for analyzing cluster-randomized trials, including the mixed-model ANCOVA, individual-level ANCOVA, and cluster-level ANCOVA estimators. These insights may inform the choice of methods in practice. Analytical results and insights are illustrated via simulation studies and analyses of three cluster-randomized trials.

C0767: Enhanced robust causal estimation of estimands in clinical trials

Presenter: **Ming Tan**, Georgetown University, United States

The central question in comparative clinical trials is how the treatment outcome compares to what would have happened to the same subjects had they received a different or no treatment, which is intrinsically a causal inference problem. In addition, after a patient is treated, intercurrent events, such as receiving rescue medicine, may depend on treatment conditions and can impact data collected to assess efficacy. The robustness of estimands of a relevant estimation, with respect to underlying assumptions, is then the key to the success of the trial, as reflected in regulatory guidance from ICH, FDA, and EMA. The aim is to present an enhanced doubly-robust estimation method utilizing semiparametric models with nonparametric monotone or concave link functions for both the propensity score and the outcome models. The models are estimated using an iterative procedure incorporating the pool adjacent violators algorithm. The asymptotic properties of the enhanced DREs are then studied. Simulation studies are performed to evaluate their finite sample performance. The benefit of this approach is explored for several causal estimands of interest. The method is then applied to analyzing several clinical trials.

C0205: **Discussion on "Statistical innovations in clinical trial design and analysis"**
Presenter: **Jeen Liu**, Regeneron Pharmaceuticals, United States

The session "Statistical innovations in clinical trial design and analysis" will be complemented by a discussion of each talk.

CO242 Room K0.20 STATISTICAL INFERENCE WITH GRAPHS Chair: Robert Lunde

C0447: **Network regression and supervised centrality estimation**
Presenter: **Ran Chen**, Washington University in St. Louis, United States
Co-authors: Junhui Cai, Haipeng Shen, Dan Yang, Linda Zhao, Wu Zhu

The centrality in a network is often used to measure nodes' importance and model network effects on a certain outcome. Empirical studies widely adopt a two-stage procedure, which first estimates the centrality from the observed noisy network and then infers the network effect from the estimated centrality, even though it lacks theoretical understanding. A unified modeling framework is proposed to study the properties of centrality estimation and inference and the subsequent network regression analysis with noisy network observations. Furthermore, a supervised centrality estimation methodology is proposed, which aims to simultaneously estimate and infer both centrality and network effect. The advantages of the method compared with the two-stage method are showcased both theoretically and numerically via extensive simulations and a case study in predicting currency risk premiums from the global trade network.

C0824: **Assumption-lean inference for the network-linked data**
Presenter: **Wei Li**, Washington University in St. Louis, United States
Co-authors: Robert Lunde, Nilanjan Chakraborty

An assumption-lean inference framework is discussed for regression models built for network-linked data. Two specific network models are explored: the graphon model and the generalized random dot product model, each with different choices of appropriate node-level network statistics. A phase transition phenomenon is discovered in both models. In denser regimes, consistent bootstrap schemes are provided for two important classes of network statistics. In sparser scenarios, a down-sampling inference method is offered that is consistent under mild conditions, albeit with a slightly slower convergence rate.

C0844: **Conformal prediction for Dyadic regression**
Presenter: **Robert Lunde**, Washington University in St Louis, United States
Co-authors: Liza Levina, Ji Zhu

Dyadic regression, which involves modeling a relational matrix given covariate information, is an important task in statistical network analysis. Uncertainty quantification is considered for dyadic regression models using conformal prediction. Novel non-conformity scores are proposed for this setting, and finite-sample validity is established in the procedures for various sampling mechanisms under a joint exchangeability assumption. It is also shown that, under certain conditions, it is possible to construct asymptotically valid prediction intervals for a missing entry under a structured missingness assumption.

C1050: **Two-sample testing with a graph-based total variation integral probability metric**
Presenter: **Alden Green**, Stanford University, United States

A novel multivariate nonparametric two-sample testing problem is considered where, under the alternative, distributions P and Q are separated in an integral probability metric over functions of bounded total variation (TV IPM). A new test, the graph TV test, is proposed, and it uses a graph-based approximation to the TV IPM as its test statistic. It is shown that this test, computed with an ε-neighborhood graph and calibrated by permutation, is minimax rate-optimal for detecting alternatives separated in the TV IPM. As an important special case, it is shown that this implies the graph TV test is optimal for detecting spatially localized alternatives, whereas the χ^2 test is probably suboptimal. The theory is supported by numerical experiments on simulated and real data.

CO269 Room K0.50 RECENT INNOVATIONS IN TWO-PHASE STUDY DESIGN AND ANALYSIS Chair: Qihuang Zhang

C0505: **Calibration methods to improve efficiency of regression analyses with two-phase samples under complex survey designs**
Presenter: **Lingxiao Wang**, University of Virginia, United States

Two-phase sampling designs are frequently employed in epidemiological studies and large-scale health surveys. In such designs, certain variables are exclusively collected within a second-phase random subsample of the initial first-phase sample, often due to factors such as high costs, response burden, or constraints on data collection or measurement assessment. Consequently, second-phase sample estimators can be inefficient due to the diminished sample size. Model-assisted calibration methods have been used to improve the efficiency of second-phase estimators. However, no existing methods provide appropriate calibration auxiliary variables while simultaneously considering the complex sample designs present in both the first- and second-phase samples in regression analyses. The proposal is to calibrate the sample weights for the second-phase subsample to the weighted entire first-phase sample based on score functions of regression coefficients by using predictions of the covariate of interest, which can be computed for the entire first-phase sample. The consistency of the proposed calibration estimation is established, and variance estimation is provided. Empirical evidence underscores the robustness of the calibration on score functions compared to the imputation method, which can be sensitive to misspecified prediction models for the variable only collected in the second phase. Examples using data from the National Health and Nutrition Examination Survey are provided.

C0527: **Studying mortality in critically ill patients: An analysis of ordinal longitudinal data under case-control sampling**
Presenter: **Chiara Di Gravio**, Imperial College London, United Kingdom
Co-authors: Ran Tao, Jonathan Schildcrout

It is common practice in clinical trials to store blood samples at recruitment and analyze them at a later stage to retrospectively obtain information on new exposures. However, high costs limit the number of samples researchers can analyze. The purpose is to discuss the experience of setting up a secondary analysis of the CLOVERS trial where, due to budget and time constraints, stored blood samples could only be analyzed on a third of the original trial population. First, the study design is introduced, and then a semiparametric likelihood approach is described to estimate the parameters of interest. The estimation approach is flexible in that it can be used for any design and exposure of interest. The finite sampling operating characteristics of the proposed approach and the results from the CLOVERS trial are finally presented.

C1041: **Design and analysis of a multi-wave two-phase study to addresses data errors in a multinational HIV research network**
Presenter: **Bryan E Shepherd**, Vanderbilt University Medical Center, United States

Routinely collected observational data from clinical and laboratory encounters are commonly used for HIV/AIDS research. There are worries about the quality of these data. The experience designing and carrying out a multi-wave validation study is described in over a dozen sites across Latin America and East Africa. The interest was to estimate the incidence of and risk factors for Kaposi Sarcoma (KS) among people living with HIV. The original error-prone dataset had data on over 257,000 patients, approximately 2,300 (<1%) with KS. A two-wave validation sample of approximately 1,000 records is designed. Optimal sampling designs that minimize the variance of resulting estimators require information that is typically not available prior to doing the data validation. Hence, approximately 500 records in a first sampling wave are validated, and this information is used to optimize the design of the second sampling wave. Finally, the analyses combined the two-waves of validation data collected in the subset of 1,000 records with the error-prone data available on the full cohort using generalized raking and multiple imputation techniques to efficiently account for the errors in the original data.

C0499: **Novel two-phase designs for evaluating new cancer screening tests**
Presenter: **Fangya Mao**, National Cancer Institute, United States

Cancer screening is evolving with advanced biomedical technologies, including novel tests for precancers and early cancers. Large-scale testing of stored study specimens to evaluate new screening tests is costly. Two-phase designs provide a robust framework for this issue. In Phase I, we gather data from the old screening test and definitive outcomes (e.g., biopsies) at study visits. Disease outcomes may be prevalent but undetected during the initial screening or incident, discovered during follow-up visits, creating left- and interval-censored time-to-event data. Phase II involves selecting a subset of subjects for the new screening test. Data analysis employs a mixture model, utilizing logistic regression to model prevalent disease risk and a proportional hazards model for incident disease risk. Various sub-sampling schemes are proposed and compared. Proposed frameworks are examined through simulations and are implemented in an evaluation of p16/ki-67 dual-stain as a triage screening test for HPV-positive women in cervical precancer screening, using stored specimens and electronic health record data from Kaiser Permanente Northern California (KPNC).

CO143 **Room K2.31 (Nash Lec. Theatre)** **NEW ADVANCES IN CAUSAL INFERENCE** **Chair: Liqun Diao**

C0398: **Optimal treatment allocation in the presence of competing risks and clustering**
Presenter: **Erica Moodie**, McGill University, Canada
Co-authors: Misha Dolmatov, Dipankar Bandyopadhyay

The precision medicine framework has been used to discover tailored treatment strategies in a variety of settings and has largely been derived within a causal framework due to the need for large (and thus typically observational) datasets. Existing methods are extended to address the pressing question of how to optimally allocate kidneys for transplantation when the pool of deceased donors includes individuals who were living with hepatitis C virus (HCV). The proposed approach accounts for the non-random allocation of kidneys from people with or without HCV, multiple (competing risks) endpoints, and clustering of data due to side effects. The proposed method is applied to data from the US National Organ Procurement and Transplant Network registry from 1994 to 2014.

C0836: **Identification and estimation of the average causal effects under dietary substitution strategies**
Presenter: **Lan Wen**, University of Waterloo, Canada

The 2020-2025 Dietary Guidelines suggest that most people can improve their diet by making some changes to what they eat and drink. In many cases, these changes involve simple substitutions. For instance, the guidelines suggest replacing processed red meat with chicken to lower sodium intake and choosing whole grains over refined grains to increase dietary fiber intake. The question about these dietary substitution strategies seeks to estimate the counterfactual mean outcome under a hypothetical intervention that replaces a food an individual would have consumed in the absence of intervention with a healthier substitute. Conditions under which the average causal effects of substitution strategies can be identified are shown. Efficient estimators for the proposed food substitution strategy are also provided, and the methodology is demonstrated via simulation studies and an application utilizing data from the Nurses Health Study.

C0908: **Nonparametric assessment of regimen response curve estimators**
Presenter: **Ashkan Ertefaie**, University of Rochester, United States
Co-authors: Cuong Pham, Benjamin Baer

Marginal structural models have been widely used in causal inference to estimate mean outcomes under either a static or a prespecified set of treatment decision rules. This approach requires imposing a working model for the mean outcome given a sequence of treatments and possibly baseline covariates. A dynamic marginal structural model is introduced that can be used to estimate an optimal decision rule within a class of parametric rules. Specifically, the mean outcome is estimated as a function of the parameters in the class of decision rules, referred to as a regimen-response curve. In general, misspecification of the working model may lead to a biased estimate with questionable causal interpretability. To mitigate this issue, the risk is leveraged to assess the goodness-of-fit of the imposed working model. The counterfactual risk is considered as the target parameter, and inverse probability weighting and canonical gradients are derived to map it to the observed data. Asymptotic properties of the resulting risk estimators are provided, considering both fixed and data-dependent target parameters. It is shown that the inverse probability weighting estimator can be efficient and asymptotic linear when the weight functions are estimated using a sieve-based estimator.

C0996: **All else being equal: Implications of measurement error for precision medicine and health equity**
Presenter: **Michael Wallace**, University of Waterloo, Canada

Precision medicine describes the tailoring of treatment decisions to individual-level characteristics. Dynamic treatment regimes operationalize precision medicine through sequences of decision rules which take patient-level data as input and output treatment recommendations. Estimation of decision rules that optimize some outcome across a population based on observational data is a large - and expanding - area of the literature. A common assumption within this framework (as well as in the broader causal inference literature) is that observed data are measured without error, which, in reality, is seldom the case. Moreover, measurement error poses some unique challenges within the context of precision medicine, such as when there is nonadherence to personalized treatment regimes or when treatment decisions are based on error-prone variates. In addition to mis-estimated optimal decision rules, a further concern arises in the context of health equity. Namely, an individual who identifies in one social (or other type of) group may be disproportionately affected if the results of an analysis based on error-prone data are implemented. The challenges are discussed and explored at the interface of precision medicine and measurement error, as well as their potential implications for health equity. In addition to theoretical results, illustrative examples are demonstrated via simulation and an R Shiny app.

CO277 **Room K2.40** **SPATIAL DATA SCIENCE** **Chair: Philipp Otto**

C0174: **Spatiotemporal comparison of sea surface to air temperatures in the tropical Pacific**
Presenter: **Peter Craigmile**, Hunter College, CUNY, United States
Co-authors: Peter Guttorp

In the study of global climate, ocean temperature estimates use sea surface temperature (SST) anomalies instead of marine atmospheric temperature (MAT) anomalies. A key question is to ask what biases result from this choice. Since SST and MAT are expected to have different correlation lengths, with SST being longer due to the slow change in ocean temperatures, it is statistically difficult to compare the two. Hierarchical statistical models are employed to investigate spatial-temporal differences between SST and MAT anomalies in the tropical Pacific. The analysis uses observations from the Tropical Atmosphere Ocean (TAO) buoy network. A spatiospectral modeling approach is used to account for missing data and quality issues in the observation network and allow for full uncertainty quantification.

C0188: **LASSO-type penalization in the framework of generalized additive models for location, scale and shape (GAMLSS)**
Presenter: **Andreas Groll**, Technical University Dortmund, Germany

A regularization approach is proposed for high dimensional data set-ups in the generalized additive model for location, scale and shape (GAMLSS) framework. It is designed for linear covariate effects and is based on L1-type penalties. The following three penalization options are provided: The conventional least absolute shrinkage and selection operator (LASSO) for metric covariates and both group and fused LASSO for categorical predictors.

C0216: **Copula-based regressions for assessing trade-offs between milk production and greenhouse-gas emissions of French farms**
Presenter: **Tristan Senga Kiesse**, Institut Agro - INRAE, France
Co-authors: Naomi Ouachene, Michael Corson, Claudia Czado

In the context of climate change, livestock systems have several challenges to meet, including improving their environmental performances and ensuring food security. Due to interactions among the many components of farms, strategies to decrease one emission may increase another emission or decrease farm production. Since the effectiveness of combined strategies to mitigate emissions at the farm scale cannot be assessed by considering the effectiveness of one strategy at a time, whole-farm mitigation scenarios are preferred. On this basis, the aim is to assess trade-offs between milk production and emissions when simultaneous changes in management practices are made. To this end, copula-based regressions are investigated to explain farm outputs, considering the multivariate dependence structure among their descriptive components. The method was applied to a dataset of management practices, production and emissions of 2523 French dairy farms surveyed in 2013. Copula-based regressions are first fitted to milk production per cow and total greenhouse gas (GHG) emissions per livestock unit separately, and then conditional dependence is assumed between them. Subsequently, we explored whole-farm mitigation scenarios to decrease GHG emissions while maintaining milk production. The utility of capturing interactions among practices and outputs of farms is then assessed in order to develop effective mitigation scenarios.

C0227: **Control charts for monitoring a BINARCH(1) process: Comparisons and applications**
Presenter: **Athanasios Rakitzis**, University of Piraeus, Greece
Co-authors: Maria Anastasopoulou

In recent years, control charts for monitoring autocorrelated count data have gained much attention, and there are many scientific works in the area. The most popular time series models of this type are those that assume first-order autocorrelation among the successive counts. The focus is on one of the most popular model in the area, the BINARCH(1) model, which is suitable for modelling first-order autocorrelated binomial counts, i.e., counts that have a bounded support. First, the available control charts in the literature are presented that can be used for the monitoring of this type of process. Then, CUSUM-type control charts are proposed and studied, which have not been studied so far. Using Monte Carlo simulation, the performance of the control charts considered is compared, either for increasing or for decreasing shifts in the process mean level. Recommendations for the most powerful scheme are also given. Finally, their practical implementation is briefly discussed via a real-data example.

CO402 Room K2.41 ROC CURVES/EVALUATION OF BIOMARKERS Chair: Vanda Inacio

C0473: **A unified framework for ROC curve inference with and without covariates**
Presenter: **Maria Xose Rodriguez Alvarez**, Universidade de Vigo, Spain

A general framework is proposed for the estimation of the receiver operating characteristic (ROC) curve and its conditional counterparts, the covariate-specific ROC curve and the covariate-adjusted ROC curve. The proposal builds upon the expression of these ROC curves as the (conditional) cumulative distribution function of the so-called (conditional) placement values, i.e., the standardization of test results in the diseased population using the non-diseased population as the reference. The validity of the approach is supported by simulations and illustrated with real data.

C0354: **Semiparametric estimator for the covariate-specific ROC curve**
Presenter: **Pablo Martinez-Camblor**, Geisel School of Medicine at Darmouth, United States
Co-authors: Juan-Carlos Pardo-Fernandez

The study of the predictive ability of a marker is mainly based on accuracy measures sures provided by the so-called confusion matrix. Besides, the area under the ROC curve, AUC, has become a popular index for summarizing the overall accuracy of a marker. However, the nature of the relationship between the marker and the outcome and the role that potential confounders play in this relationship could be fundamental in extrapolating the observed results. Directed acyclic graphs (DAGs), commonly used in epidemiology and causality, could provide good feedback for learning the possibilities and limits of this extrapolation applied to the binary classification problem. Both the covariate-specific and the covariate-adjusted ROC curves are valuable tools that can help to better understand the real classification abilities of a marker. Since they are strongly related to the conditional distributions of the marker on the positive (subjects with the studied characteristic) and negative (subjects without the studied characteristic) populations, the use of proportional hazard regression models arises in a very natural way. The use of flexible proportional hazard Cox regression models is explored for estimating the covariate-specific and the covariate-adjusted ROC curves. Their large- and finite-sample properties are studied, and the proposed estimators are applied to a real-world problem.

C0939: **Evaluating prognostic biomarkers for censored survival data with covariate adjustment**
Presenter: **Ainesh Sewak**, University of Zurich, Switzerland
Co-authors: Vanda Inacio, Torsten Hothorn

Identifying reliable biomarkers for predicting clinical events in longitudinal studies is important for enabling accurate disease prognosis and supporting the development of new therapies. Traditional receiver operating characteristic (ROC) curve analysis has been adapted for time-dependent and censored outcomes. However, accounting for clinical heterogeneity in patient characteristics remains a challenge in assessing the prognostic accuracy of biomarkers. Prior methods have relied on the proportional hazards assumption or model only the summary statistics of the ROC curve. A flexible conditional bivariate model is proposed to quantify biomarkers' prognostic accuracy for censored survival data while accounting for covariates. The model uses separate marginal regression models for the time-to-event and biomarker outcomes, accounting for their dependence structure on a latent normal transformed scale. By parameterizing the marginal models, the maximum likelihood for estimation and inference is used. The application of the method in a study is demonstrated by examining serum neurofilament biomarkers for predicting survival in amyotrophic lateral sclerosis (ALS) patients.

C0923: **Biomarker cutoff estimation and their confidence intervals under ternary umbrella and tree stochastic ordering settings**
Presenter: **Leonidas Bantis**, University of Kansas Medical Center, United States
Co-authors: Benjamin Brewer

Tuberculosis (TB) studies often involve four different states under consideration, namely, healthy, latent infection, pulmonary active disease, and extra-pulmonary active disease. While diagnostic tests do exist, they are expensive and generally not accessible in regions where they are most needed; thus, there is an interest in assessing the accuracy of new and easily obtainable biomarkers. For some such biomarkers, the typical stochastic ordering assumption might not be justified for all disease classes under study, and usual ROC methodologies that involve ROC surfaces and hypersurfaces are inadequate. Different types of orderings may be appropriate depending on the setting, and these may involve a number of ambiguously ordered groups that stochastically exhibit larger (or lower) marker scores than the remaining groups. Recently, there has been scientific interest in ROC methods that can accommodate these so-called tree or umbrella orderings. However, there is limited work discussing the estimation of cutoffs in such settings. The purpose is to discuss the estimation and inference around optimized cutoffs when accounting for such configurations. Different cutoff alternatives are explored, and parametric is provided, flexible parametric and non-parametric kernel-based approaches for estimation and inference. The approaches are evaluated using simulations and are illustrated through a real data set that involves TB patients.

CO164 Room S0.03 HANDLING COMPLEX DATA AND COMPLEXITY Chair: Jacopo Di Iorio

C1039: Feature generating models: Inference in purely high dimensions

Presenter: **Benjamin Roycraft**, University of Florida, United States

The significance of high-dimensional data lies in its pervasive presence across numerous scientific, engineering, and business domains. As datasets grow in complexity and scale, the analysis of high-dimensional data becomes increasingly vital. In fields like genomics, health sciences, and finance, where intricate relationships and interactions abound, the ability to navigate and derive meaningful insights from large datasets is crucial. Whether it's understanding protein interactions, optimizing financial portfolios with thousands of assets, or interpreting high-resolution images, the capacity to handle and analyze data with a large number of correlated variables is at the forefront of advancements in research, technology, and innovation. A new modelling framework is presented, which allows for inference, variable selection, and dimension reduction in the most challenging purely-dimensional asymptotic regime, where the sample size is fixed, and the number of observed variables grows without bounds.

C1377: Modeling spatial anisotropy and non-stationarity in semiparametric regression with differential penalization

Presenter: **Eleonora Arnone**, University of Turin, Italy
Co-authors: Matteo Tomasetto, Laura Sangalli

Spatial regression models are essential for analyzing environmental data and predicting phenomena that vary across space. However, conventional methods often fall short of capturing the intricate spatial dependencies and non-stationarities found in real-world datasets. The purpose is to introduce a novel parameter-cascading algorithm for spatial regression. This algorithm simultaneously estimates the unknown spatial parameters that describe anisotropy and the spatial field itself while integrating physical and domain-specific knowledge. The efficacy of the parameter-cascading algorithm is demonstrated through simulation studies and by applying it to a case study involving environmental data. Through this application, the method is shown to improve the accuracy of predictions for spatially distributed variables.

C1591: Forecasting curves portions using motif discovery inspired method

Presenter: **Yijiang Fan**, Sant Anna School of Advanced Studies, Italy

A nonparametric method for forecasting in functional data analysis is proposed. We address forecasting the last portion of curves, which can be extended to the imputation of missing portions in the curves. The forecast method is based on the notion of functional motifs, which are patterns that recur in multiple portions of a single curve or in multiple curves. Taking the last portion of a curve as a segment of a candidate motif, the other occurrences of the candidate motif are found; the forecast is the forward projection of the recurrent motif. The feasibility of the proposed forecast method is assessed with diagnostic methods that evaluate whether the last portion of a curve is an occurrence of a motif or not. The performance of the method is examined through simulations of multiple scenarios compared with benchmark methods in functional data forecasting. Eventually, the method is applied to real-world climate data.

C1610: Distilling causal effects: Stable subgroup estimation via distillation trees in causal inference

Presenter: **Ana Kenney**, University of California, Irvine, United States
Co-authors: Tiffany Tang, Melody Huang

Researchers are interested in understanding the underlying treatment effect heterogeneity. While recent methodological developments have introduced new black-box approaches to better estimate heterogenous treatment effects, these methods only provide an estimate of the individual-level treatment effect and fall short of characterizing the underlying individuals who may be most at risk or benefit most from receiving the treatment. A method, causal distillation trees (CDT), is introduced that allows researchers to estimate interpretable subgroups in their studies stably. CDT allows researchers to fit any machine learning model of their choice to estimate the individual-level treatment effect and then leverages a simple, second-stage tree-based model to distil the estimated treatment effect into meaningful subgroups. As a result, CDT inherits the theoretical guarantees from black-box machine learning models, while preserving the interpretability of a simple decision tree. The stability of CDT is theoretically characterized by estimating substantively meaningful subgroups, and helpful diagnostics are provided for researchers to evaluate the quality of the estimated subgroups. The method is empirically demonstrated via extensive simulations and a case study evaluating the impact of canvassing on voter turnout. It is shown that CDT out-performs state-of-the-art approaches in identifying interpretable subgroups.

CO067 Room S0.11 NOVEL STATISTICAL METHODS AND ANALYSES FOR NEUROIMAGING AND BIOMEDICAL DATA Chair: Cai Li

C1274: Sex-specific topological structure associated with dementia via latent space estimation

Presenter: **Selena Wang**, Indiana University School of Medicine, United States

Sex-specific topological structure associated with typical Alzheimer's disease (AD) dementia is investigated using a novel state-of-the-art latent space estimation technique. A probabilistic approach for latent space estimation extends current multiplex network modeling approaches and captures the higher-order dependence in functional connectomes by preserving transitivity and modularity structures. Sex differences are found in network topology, with females showing more default mode network (DMN)-centered hyperactivity, whereas males show more limbic system (LS)-centered hyperactivity while both show DMN-centered hypoactivity. Centrality plays are found to have an important role in dementia-related dysfunction, with a stronger association between connectivity changes and regional centrality in females than in males. The contribution to current literature is that it provides a more comprehensive picture of dementia-related neurodegeneration linking centrality, network segregation, and DMN-centered changes in functional connectomes and how these components of neurodegeneration differ between the sexes.

C1277: Fiber tract microstructural quantile (FMQ) regression for white matter tracts

Presenter: **Zhou Lan**, Brigham and Womenś Hospital, Harvard Medical School, United States

The brain's white matter is critical for cognition. A new approach for statistical analysis of white matter fiber tracts, fiber tract microstructural quantile (FMQ) regression, is introduced. The method employs the statistical technique of quantile regression with clustered data to investigate the relationship between fiber tract tissue microstructure and clinical or psychological covariates. To demonstrate the proposed approach, an illustrative study is provided based on the data of a large dataset, human connectome project-young adult (HCP-YA), with a focus on specific tracts expected to relate to particular aspects of motor function and cognition as described in a recent review. The cohort of 809 participants is used for the illustrative study. The arcuate fasciculus (AF), uncinate fasciculus (UF), Cingulum (CB), and corticospinal tract (CST) were selected to investigate their associations with scalar factors of language, memory, executive function, and motor, respectively. The illustrative study results follow the previously established findings and imply that our proposed profile is more powerful in identifying significance than the methods compared. Moreover, it is demonstrated that using the quantile profile might be more anatomically insightful in providing microstructural inference for investigating the association between scalar factors and white matter tracts.

C1331: **Imaging mediation analysis for longitudinal outcomes**
Presenter: **Cai Li**, St. Jude Children's Research Hospital, United States

The focus is on improving cognitive outcomes for pediatric cancer survivors who undergo aggressive cancer treatments that may affect the central nervous system. Specifically, a new mediation framework is proposed for longitudinal neurocognitive outcomes pertaining to a clinical trial for medulloblastoma, the most common malignant brain tumour in children, using high-dimensional imaging mediators to identify causal pathways and corresponding white matter microstructures. The proposed approach takes into account both the spatial and temporal dependencies and smoothness of the mediators and outcomes, enhancing the detection power of informative voxels and accurately characterizing longitudinal patterns concurrently. The results offer insights into how to enhance long-term neurodevelopment and strategically spare brain regions that might be impacted by radiation therapy. This understanding will be crucial in planning future treatment protocols, ultimately benefiting brain cancer survivors.

C1567: **Efficient fully Bayesian approach to brain activity mapping with complex-valued fMRI data**
Presenter: **Andrew Brown**, Clemson University, United States

Functional magnetic resonance imaging (fMRI) enables indirect detection of brain activity changes via the blood-oxygen-level-dependent (BOLD) signal. Conventional analysis methods mainly rely on the real-valued magnitude of these signals. In contrast, research suggests that analyzing both real and imaginary components of the complex-valued fMRI (cv-fMRI) signal provides a more holistic approach that can increase power to detect neuronal activation. A fully Bayesian model for brain activity mapping is proposed with cv-fMRI data. The model accommodates temporal and spatial dynamics. Additionally, a computationally efficient sampling algorithm is proposed, which enhances processing speed through image partitioning. The approach is shown to be computationally efficient via image partitioning and parallel computation while being competitive with state-of-the-art methods. These claims are supported by both simulated numerical studies and an application to real cv-fMRI data obtained from a finger-tapping experiment.

CO265 **Room S0.12** **METHODS FOR MULTIPARAMETER EVIDENCE SYNTHESIS AND SPATIAL OMICS DATA** **Chair: Zelalem Negeri**

C0284: **Advancing trial sequential analyses for living systematic reviews**
Presenter: **Lifeng Lin**, University of Arizona, United States

A living systematic review (LSR) is a progressive approach aimed at providing ongoing updates and instant synthesis of evidence. Trial sequential analysis (TSA) is an important tool for evaluating the sufficiency of evidence gathered in an LSR. It utilizes trial sequential monitoring boundaries to assess the effectiveness of an intervention and futility boundaries to determine if the intervention does not significantly differ from the control. While TSAs have been increasingly popular, their reproducibility is currently limited due to a lack of detailed information on their assumptions. Moreover, existing TSA methods face challenges due to their significant reliance on interim analyses of randomized controlled trials, which typically involve more homogeneous participant groups than those found in meta-analyses. New methods are introduced aimed at preventing premature terminations of LSRs, thereby enabling more robust evidence syntheses. Numerical studies indicate that these proposed methods are more reliable than current methods.

C0760: **Comparison of dose-response meta-analytic models using empirical and simulation studies**
Presenter: **Joseph Beyene**, McMaster University, Canada

Dose-response relationship studies are common in clinical as well as epidemiological studies and are critical for understanding how varying levels of exposure impact the risk of an outcome. Using meta-analytic approaches to synthesize dose-response data from multiple studies presents several challenges. One-stage and two-stage dose-response meta-analysis (DRMA) methods are commonly employed. The one-stage method employs a linear mixed model, while the two-stage method involves estimating key model parameters within each study and synthesizing them across studies. The underlying dose-response relationship can be linear or nonlinear, depending on the specific exposure-outcome pairing. Various exposure-outcome relationships are investigated using the one-stage DRMA method, employing linear, quadratic polynomial, and restricted cubic spline (RCS) models. Knot selection in RCS models, model selection using different functional forms, and the influence of outlying studies are investigated in detail. Various models are assessed using an empirical study based on a large collection of published DRMA datasets. With model parameter selection informed by the empirical data, extensive simulations are designed and implemented to evaluate comparative performance across several realistic scenarios.

C0682: **Precision of treatment hierarchy: A metric for quantifying certainty in treatment hierarchies from network meta-analysis**
Presenter: **Augustine Wigle**, University of Waterloo, Canada
Co-authors: Audrey Beliveau, Georgia Salanti, Gerta Rucker, Guido Schwarzer, Dimitris Mavridis, Adriani Nikolakopoulou

Network meta-analysis (NMA) is an extension of pairwise meta-analysis which facilitates the estimation of relative effects for multiple competing treatments. A hierarchy of treatments is a useful output of an NMA. Treatment hierarchies are produced using ranking metrics. Common ranking metrics include the Surface Under the Cumulative RAnking curve (SUCRA) and P-scores, which are the frequentist analogue to SUCRAs. Both metrics consider the size and uncertainty of the estimated treatment effects, with larger values indicating a more preferred treatment. Although SUCRAs and P-scores themselves consider uncertainty, treatment hierarchies produced by these ranking metrics are typically reported without a measure of certainty, which might be misleading to practitioners. We propose a new metric, Precision of Treatment Hierarchy (POTH), which quantifies the certainty of producing a treatment hierarchy from SUCRAs or P-scores. POTH provides a single, interpretable value which quantifies the degree of certainty in producing a treatment hierarchy. We show how the metric can be adapted to apply to subsets of treatments in a network, for example, to quantify the certainty in the hierarchy of the top three treatments. We calculate POTH for a database of NMAs to investigate its empirical properties, and we demonstrate its use on three published networks.

C0737: **Multi-parameter estimation of prevalence (MPEP) models to estimate the prevalence of opioid dependence**
Presenter: **Hayley Jones**, University of Bristol, United Kingdom
Co-authors: Matthew Hickman, Andreas Markoulidakis

Estimates of the number of people dependent on opioids are critical for planning public health services and evaluating interventions to reduce drug-related harms. So-called indirect methods are needed since population surveys underestimate the extent of stigmatized behaviors. The MPEP approach is developed to estimate prevalence or population size. MPEP utilizes routinely collected linked administrative data on opioid agonist therapy (OAT) prescriptions and adverse events such as opioid-related deaths or non-fatal overdoses. The known population of people with opioid dependence is identified as everyone who recently received OAT, and the size of the additional unknown population is estimated through modelling of adverse event rate data, critically including events occurring outside of the known population. In a joint model fitted in a Bayesian framework, simultaneous regressions are fitted to event rates and (latent) prevalence. Importantly, the impact of OAT on event risk is accounted for since everyone in the unknown group is not receiving OAT, by definition. Joint modelling of two or more types of adverse events allows checking the consistency of evidence. The MPEP approach, its assumptions and limitations, and the recent application are described to estimate the prevalence of opioid dependence in Scotland from 2014/15 to 2019/20.

C0726: **Enhance constraint spatial partitioning for spatial omics data**
Presenter: **Pratheepa Jeganathan**, McMaster University, Canada
Co-authors: Rajitha Senanayake

Spatial omics data, when preprocessed, identifies cell types across multiple tissues, providing crucial insights into the tumor microenvironment

(TME). Traditional methods, such as constraint spatial hierarchical clustering and spatial latent Dirichlet allocation (sLDA), have characterized TME by classifying it in patients and associating it with clinical factors like survival time and treatment type. However, a significant limitation of these methods is their inability to uncover spatially separated partitions with similar cell-type distributions. To address this limitation, the enhanced constraint spatial partitioning (ECSP) approach is proposed, which leverages a soft constraint hierarchical clustering framework to merge similar partitions in cell-type distribution. Additionally, sLDA is demonstrated to be improved by incorporating a relational topic model framework to generate links after identifying cell partitions. The methods enhance the interpretability of the TME by effectively integrating spatial context. Their effectiveness is demonstrated using multiplexed ion beam imaging-time of flight (MIBI-TOF) data, highlighting their potential to improve TME characterization and clinical associations. These novel deterministic and probabilistic approaches promise to refine the analysis of spatial omics data, providing a deeper understanding of the spatial distribution and interaction of cell types within the tissues.

CO029 Room S0.13 RECENT ADVANCES IN COMPLEX ANALYSIS OF GENOMICS DATA Chair: Li-Xuan Qin

C0311: **Advances in identifying latent gene-gene and gene-environment interactions for binary outcomes**
Presenter: **Lei Sun**, University of Toronto, Canada

Investigating gene-environment (GxE) interactions without direct environmental data poses challenges, and exhaustive gene-gene (GxG) searches face issues with large-scale multiple-hypothesis testing. These complexities are explored, focusing on binary traits. For continuous traits, latent interactions induce artificial heteroskedasticity, which is detectable with methods like Levene's test. However, binary traits present unique challenges due to their variance being determined by the mean. These challenges are addressed, and a novel approach is proposed and evaluated based on non-additive genetic effects. Practical insights are demonstrated through an application to the UK Biobank data.

C1024: **Novel mediation analysis with high-dimensional omics mediators**
Presenter: **Peng Wei**, The University of Texas MD Anderson Cancer Center, United States

Environmental exposures can regulate intermediate molecular phenotypes, such as the transcriptome, metabolome, and microbiome, by various mechanisms and thereby lead to different health outcomes. It is of significant scientific interest to unravel the role of potentially high-dimensional intermediate phenotypes in the relationship between environmental exposure and health traits. Mediation analysis is an important tool for investigating such relationships. However, there are many unique challenges facing high-dimensional mediation analysis with these emerging omics mediators. To this end, an R-squared (R2) total mediation effect size measure is extended for continuous outcomes, originally proposed in the single-mediator setting, to the moderate- and high-dimensional mediator settings in the mixed model framework. Some recent advances are introduced in R2-based mediation analysis with high-dimensional omics mediators, including speeding up confidence interval estimation based on asymptotic results, extension to time-to-event and binary outcomes, meta-analysis, and applications to the Trans-Omics for Precision Medicine (TOPMed) program cohorts.

C1025: **Response variable selection in multivariate linear regression**
Presenter: **Kshitij Khare**, University of Florida, United States

In some applications involving multivariate linear regression, it is of scientific interest to identify/select responses that have at least one nonzero regression coefficient. These are referred to as dynamic responses. Because of the asymmetric roles of the predictors and responses in regression, response variable selection is markedly different from the usual predictor variable selection. In particular, when a response is inferred to have all regression coefficients equal to zero, it should not be simply removed from subsequent estimation. If it is correlated with the dynamic responses given all other responses, it should be retained to improve estimation efficiency as an ancillary statistic. Otherwise, it can be removed from further inference, and it is called a static response. Therefore, the responses can be classified into three categories: dynamic responses, ancillary responses, and static responses. An algorithm is derived to identify these response variables, and an estimator of the regression coefficients is provided based on the selection result. The scientific insights and efficiency gains obtained by the proposed procedure are illustrated with data. Consistency of the selection procedures and asymptotic properties of the estimators are established both for the large sample size and the high-dimensional small sample size settings.

C1021: **Optimizing sample size for statistical learning with bulk transcriptomic sequencing: A learning curve approach**
Presenter: **Li-Xuan Qin**, Memorial Sloan Kettering Cancer Center, United States
Co-authors: Yunhui Qi

Accurate sample classification using transcriptomics data is crucial for personalized medicine. The success of such endeavors depends on determining a suitable sample size and ensuring adequate statistical power without unnecessary resource allocation or ethical concerns. Current sample size calculation methods for sample classification rely on assumptions and algorithms that may not align with modern machine-learning techniques. The methodological gap is addressed by developing computational approaches to determine the required number of samples for accurate classification in transcriptomics studies using machine learning. The approach establishes the power-versus-sample-size relationship by employing a data augmentation strategy followed by fitting a learning curve. Its performance is evaluated for both microRNA and RNA sequencing using data from the Cancer Genome Atlas, considering various data characteristics (such as sample size, marker filtering, and sequencing depth normalization) and algorithm configurations (including model selection, hyperparameter tuning, and offline augmentation), based on a range of evaluation metrics. Python and R code for implementation of the proposed approach is freely available on GitHub. The adoption of machine learning in biomedical transcriptomics studies is expected to advance and accelerate their translation into clinically useful classifiers.

CO034 Room Safra Lec. Theatre STATISTICAL LEARNING WITH COMPLEX FUNCTIONAL DATA Chair: Haolun Shi

C0414: **Identification of regions of interest in neuroimaging data based on semiparametric transformation models**
Presenter: **Haolun Shi**, Simon Fraser University, Canada

Alzheimer's disease (AD) is a progressive neurodegenerative disorder that leads to memory loss, cognitive decline, and behavioral changes without a known cure. Neuroimages are often collected alongside the covariates at baseline to forecast the prognosis of the patients. Identifying regions of interest within the neuroimages associated with disease progression is thus of significant clinical importance. One major complication in such analysis is that the domain of the brain area in neuroimages is irregular. Another complication is that the time-to-AD is interval-censored, as the event can only be observed between two revisit time points. To address these complications, the proposal is to model the imaging predictors via bivariate splines over triangulation and incorporate the imaging predictors in a flexible class of semiparametric transformation models. The regions of interest can then be identified by maximizing a penalized likelihood. A computationally efficient expectation-maximization algorithm is devised for parameter estimation. An extensive simulation study is conducted to evaluate the finite-sample performance of the proposed method. An illustration with the Alzheimer's disease neuroimaging initiative dataset is provided.

C1584: **Representation learning of dynamic networks**
Presenter: **Haixu Wang**, University of Calgary, Canada

A representation learning model for dynamic networks, which describes the continuously changing relationship between individuals in a population, is established. The problem is encapsulated in a dimension reduction problem of functional data analysis. Given that the dynamic networks are a form of matrix-valued functions, the goal is to find a mapping that compresses such functional data into vector-valued functions. The learning space, which is a much lower dimensional function space, is endowed with norm and inner product. Moreover, the learning method is further allowed to be asymmetric in the learning space so that the in and out roles of individual nodes can be studied separately. In simulation studies, the finite sample performance of the method in link prediction is compared to the existing methods. Sociological networks are common examples in real data applications. The representation learning method is applied to the social networks of six ant colonies where network connections are observed over a time span of 41 days. The interactions with colonies are well captured in the low-dimensional

C1592: **Partial least squares for sparsely observed curves with measurement errors**
Presenter: **Zhiyang Zhou**, University of Wisconsin-Milwaukee, United States

Functional partial least squares (FPLS) are commonly used for fitting scalar-on-function regression models. For the sake of accuracy, FPLS demands that each realization of the functional predictor be recorded as densely as possible over the entire time span; however, this condition is sometimes violated in longitudinal studies and missing data research. Targeting this point, FPLS is adapted to scenarios in which the number of measurements per subject is small and bounded from above. The resulting proposal is abbreviated as PLEASS. Under certain regularity conditions, the consistency of estimators is established, and confidence intervals are given for scalar responses. Simulation studies and real-data applications illustrate the competitive accuracy of PLEASS.

C1628: **Function-on-function combined regression models**
Presenter: **Tianyu Guan**, York University, Canada
Co-authors: Shifan Jia, Haolun Shi

A function-on-function combined regression model is proposed that predicts a functional response by both a nonlinear dynamic effect of a functional predictor and a linear concurrent effect of another functional predictor. The nonlinear dynamic effect is characterized by taking an integral of a time-dependent two-dimensional smooth surface, and the linear concurrent effect is modeled through a time-varying coefficient. The model structure combines the flexibility of nonlinear modeling with the interpretability of the linear concurrent effect. To approximate the two-dimensional surface, tensor product basis expansions are used, and for the time-varying coefficient in the concurrent effect, B-spline expansions are employed. The expansion parameters for each effect are estimated iteratively to account for the mutual dependencies between these two estimated effects. Each iteration of parameter estimation involves solving a penalized least squares problem. The asymptotic properties of the estimator are established. The numerical performance of the proposed method is illustrated by simulation studies and two real data applications.

CO252 Room BH (S) 1.01 Lec. Theatre 1 MODERN CHANGE-POINT ANALYSIS Chair: Hao Chen

C0241: **Online kernel CUSUM for change-point detection**
Presenter: **Yao Xie**, Georgia Institute of Technology, United States
Co-authors: Song Wei

A computationally efficient online kernel cumulative sum (CUSUM) method is presented for change-point detection that utilizes the maximum over a set of kernel statistics to account for the unknown change-point location. The approach exhibits increased sensitivity to small changes compared to existing kernel-based change-point detection methods, including scan-B statistic, corresponding to a non-parametric Shewhart chart-type procedure. Accurate analytic approximations are provided for two key performance metrics: the average run length (ARL) and expected detection delay (EDD), which enable establishing an optimal window length to be on the order of the logarithm of ARL to ensure minimal power loss relative to an oracle procedure with infinite memory. Moreover, a recursive calculation procedure is introduced for detection statistics to ensure constant computational and memory complexity, which is essential for online implementation. Through extensive experiments on both simulated and real data, the competitive performance of the method is demonstrated, and the theoretical results are validated.

C0246: **Segmenting watermarked texts from language models**
Presenter: **Xianyang Zhang**, Texas A&M University, United States
Co-authors: Guanxun Li, Xingchi Li

Watermarking is a technique that involves embedding (nearly) unnoticeable statistical signals within generated content to help trace its source. The focus is on a scenario where an untrusted third-party user sends prompts to a trusted language model (LLM) provider, who then generates a text from their LLM with a watermark. This setup makes it possible for a detector to later identify the source of the text if the user publishes it. The user can modify the generated text by substitutions, insertions, or deletions. The objective is to develop a statistical method to detect if a published text is LLM-generated, mainly from the perspective of a detector. A methodology is further proposed to segment the published text into watermarked and non-watermarked sub-strings. The proposed approach is built upon randomization tests and change point detection techniques. The demonstrated method ensures Type I and Type II error control and can accurately identify watermarked sub-strings by finding the corresponding change point locations. To validate the technique, it is applied to texts generated by several language models with prompts extracted from Google's C4 dataset and obtain encouraging numerical results.

C0730: **Asymptotic distribution-free change-point detection for modern data based on a new ranking scheme**
Presenter: **Doudou Zhou**, National University of Singapore, Singapore
Co-authors: Hao Chen

Change-point detection (CPD) involves identifying distributional changes in a sequence of independent observations. Among nonparametric methods, rank-based methods are attractive due to their robustness and effectiveness and have been extensively studied for univariate data. However, they are not well explored for high-dimensional or non-Euclidean data. A new method, rank induced by graph change-point detection (RING-CPD), is proposed, which utilizes graph-induced ranks to handle high-dimensional and non-Euclidean data. The new method is asymptotically distribution-free under the null hypothesis, and an analytic p-value approximation is provided for easy type-I error control. Simulation studies show that RING-CPD effectively detects change points across a wide range of alternatives and is also robust to heavy-tailed distribution and outliers. The new method is illustrated by the detection of seizures in a functional connectivity network dataset, changes in digit images, and travel pattern changes in the New York City Taxi dataset.

C0816: **Two-stage sequential change diagnosis problems**
Presenter: **Lifeng Lai**, University of California, Davis, United States

A two-stage Bayesian sequential change diagnosis (SCD) problem is formulated and solved. In the SCD problem, after a change is detected, one also needs to determine what the post-change distribution is. Different from the one-stage sequential change diagnosis problem considered in the existing work, after a change has been detected, it is possible to continue collecting low-cost samples so that the post-change distribution can be identified more accurately. The goal of a two-stage SCD rule is to minimize the total cost, including delay, false alarm probability, and misdiagnosis probability. To solve the two-stage SCD problem, the problem is first converted into a two-ordered optimal stopping time problem. Using tools from optimal multiple stopping time theory, the optimal SCD rule is obtained. Moreover, to address the high computational complexity issue of the optimal SCD rule, a computationally efficient threshold-based two-stage SCD rule is further proposed. By analyzing the asymptotic behaviors of the delay, false alarm, and misdiagnosis costs, it is shown that the proposed threshold SCD rule is asymptotically optimal as the per-unit delay costs go to zero.

CO120 Room BH (SE) 1.01 BAYESIAN STATISTICAL LEARNING IN PRACTICE Chair: Alejandro Murua

C0352: **Performance of the annealed MALA in transience and in stationarity**
Presenter: **Mylene Bedard**, University of Montreal, Canada

Statistical models have been increasing both in terms of complexity and dimensionality. These models cannot be treated analytically; Markov chain Monte Carlo (MCMC) methods have thus become a device of choice to sample from such target distributions. A generalized version of the Metropolis-adjusted Langevin algorithm (MALA) is introduced that features two tuning parameters: the usual step size and an interpolation parameter that accommodates the dimension of the target distribution. The efficiency of this sampler is theoretically studied by making use of the local- and global-balance concepts of a prior study, and user-friendly tuning guidelines are provided. Although the traditional MALA is theoretically optimal in infinite-dimensional settings, in practice, the annealed MALA remains superior in all contexts. It offers significant efficiency gains both in transience and in stationarity at no extra computational cost. Simulation studies corroborate the findings. In particular, the efficiency of the annealed MALA compares favorably to that of competing algorithms in various Bayesian logistic regression contexts.

C0751: **Incorporating gene ontology and disease ontology into Bayesian feature selection for cancer subtypes**

Presenter: **Thierry Chekouo**, University of Minnesota, United States

Kidney and lung cancers are among the deadliest diseases and have multiple subtypes. A Bayesian variable selection method is proposed for genomic selection and survival prediction, incorporating gene ontology and disease ontology information with application to the cancer genome atlas's kidney and lung cancer repositories. Instead of treating each gene equally without regard to their molecular functions as in conventional Bayesian variable selection methods, an algorithm and a Markov chain-like prior are proposed for the inclusion probabilities of genes that account for the functional similarities of genes. The aim is to select genes with functional similarity, even if they are highly correlated or coexpressed. Also, as a pan-cancer model, the regression model of different tumor types is linked so that the tumor types with small sample sizes can borrow information from other regressions and the statistical power of regressions with small sample sizes can be better enhanced. A prior is proposed to correlate the inclusion statuses of the same gene in different regressions, and the strength of correlation depends on the disease ontology semantic similarity between corresponding tumor types. The proposed method can address some obstacles facing other pan-cancer models with improved variable selection and prediction performances, as demonstrated in the simulation studies.

C0889: **High-dimensional variable selection in non-linear mixed-effects models**

Presenter: **Guillaume Kon Kam King**, Universita Paris-Saclay, INRAE, France

High-dimensional data (many more covariates than observations) are now commonly analyzed. However, there are few tools for high-dimensional variable selection when the data are observations collected repeatedly on several individuals, and even fewer when the model is nonlinear. Thus, a high-dimensional covariate selection method is developed for nonlinear mixed-effects models, which are natural tools for analyzing data of this nature. More precisely, it is proposed to use a spike and slab prior for variable selection, coupled with a stochastic approximation version of the EM algorithm for scalability. Targeting the maximum a posteriori, the proposed approach is much faster than classical MCMC procedures and shows very good selection performance on simulated data.

C1350: **Statistical inference for privatized data with unknown sample size**

Presenter: **Andres Barrientos**, Florida State University, United States

Co-authors: Jordan Awan, Nianqiao Ju

Theory and algorithms are developed to analyze privatized data in unbounded differential privacy(DP), where even the sample size is considered a sensitive quantity that requires privacy protection. It is shown that the distance between the sampling distributions under unbounded DP and bounded DP goes to zero as the sample size n goes to infinity, provided that the noise used to privatize n is at an appropriate rate; ABC-type posterior distributions are also established to converge under similar assumptions. Asymptotic results are further given in the regime where the privacy budget for n goes to zero, establishing the similarity of sampling distributions as well as showing that the MLE in the unbounded setting converges to the bounded-DP MLE. In order to facilitate valid, finite-sample Bayesian inference on privatized data in the unbounded DP setting, a reversible jump MCMC algorithm is proposed, which extends the data augmentation MCMC of another study. A Monte Carlo EM algorithm is also proposed to compute the MLE from privatized data in both bounded and unbounded DP. The focus is on describing these results and discussing how they can be used to select an appropriate differentially private framework for validation servers. A validation server is a secure system that allows users to query data while ensuring the privacy of the data subjects by only returning results that adhere to specified privacy constraints.

CO181 Room BH (SE) 1.02 ADVANCES IN BAYESIAN SMOOTHING AND RECURSIVE PARTITIONING Chair: Sameer Deshpande

C0298: **Scalable targeted smoothing in high-dimensions with BART**

Presenter: **Sameer Deshpande**, University of Wisconsin–Madison, United States

Bayesian additive regression trees (BART) is an easy-to-use and highly effective nonparametric regression model that approximates unknown functions with a sum of binary regression trees (i.e., piecewise-constant step functions). Consequently, BART is fundamentally limited in its ability to estimate smooth functions. Initial attempts to overcome this limitation replaced the constant output in each leaf of a tree with a realization of a Gaussian process (GP). While these elaborations are conceptually elegant, most implementations thereof are computationally prohibitive, displaying a nearly-cubic per-iteration complexity. A version of BART is proposed, built with trees that output linear combinations of ridge functions; that is, the trees return linear combinations of compositions between affine transforms of the inputs and a (potentially non-linear) activation function. A new MCMC sampler is developed to update trees in linear time. The proposed model includes a random Fourier feature-inspired approximation to treed GPs as a special case. More generally, the proposed model can be viewed as an ensemble of local neural networks, which combines the representational flexibility of neural networks with the uncertainty quantification and computational tractability of BART.

C0326: **Geometric shapes of the tree-induced partition**

Presenter: **Hengrui Luo**, Rice University, United States

Decision trees are a crucial class of regression methods in modern statistics and machine learning. Traditionally, these methods create partitions in the form of nested rectangles. However, real-world applications often demand more flexible and irregular partition shapes. The computational complexity and scalability of decision trees when dealing with such irregular geometric partitions are explored. It demonstrates how decision trees effectively create multidimensional boundaries to divide the input space, highlighting the increased search difficulty and computational challenges that arise. To address these complexities, an innovative tensor-on-tensor tree regression approach is introduced, specifically designed to handle multidimensional geometric partitioning. Finally, the fundamental geometric principles underlying tree-induced partitions are reflected on, and future research directions at the intersection of geometry and tree-based models are considered.

C0416: **Residual tree Gaussian processes**

Presenter: **Pulong Ma**, Iowa State University, United States

Gaussian processes (GP) enjoy wide popularity in spatial statistics, uncertainty quantification, and machine learning. With the advance of measurement technologies and increasing computing power, large numbers of measurements and large-scale numerical simulations make traditional GP models and computational strategies inadequate in dealing with spatially heterogeneous and big data, especially in multi-dimensional domains. In recent years, several multi-scale or tree-based extensions of the GP have been introduced to model spatial nonstationarity and/or achieve scalable computation. A new Bayesian tree-based GP inference framework, called residual treed GP (ResTGP), is introduced. ResTGP integrates the divide-and-conquer and the multi-scale modeling strategies, thereby enjoying the computational efficiency of the formal and the flexibility of the latter. The main idea is to decompose a Gaussian process as well as the data at a cascade of resolutions across locations through iteratively computing predictive and residual processes, thereby characterizing the underlying covariance structure and achieving divide-and-conquer on the data points simultaneously. A new computational strategy is also introduced for Bayesian inference for ResTGP that does not rely on Metropolis-Hastings-based stochastic tree search algorithms but is based on recursive message passing.

C0948: **Efficient non-Gaussian variational inference for continuous functions using sparse autoregressive normalizing flows**
Presenter: **Paul Wiemann**, The Ohio State University, United States
Co-authors: Matthias Katzfuss

An innovative framework is proposed for efficient and flexible variational inference aimed at non-Gaussian posteriors of latent continuous functions or fields. The framework employs sparse autoregressive structures represented by nearest-neighbor-directed acyclic graphs for both the prior and variational families. The conditional distributions within the variational family are modeled using normalizing flows, providing high flexibility. An algorithm is introduced for doubly stochastic variational optimization, achieving polylogarithmic time complexity per iteration. Preliminary numerical comparisons suggest that the proposed method may surpass the accuracy of Gaussian variational families while maintaining comparable computational complexity.

CO186 **Room BH (SE) 1.06** **DISTRIBUTION-LEAN STATISTICAL INFERENCE** **Chair: Arun Kuchibhotla**

C1502: **Inference for median and a generalization of HulC**
Presenter: **Manit Paul**, Wharton School, University of Pennsylvania, United States
Co-authors: Arun Kuchibhotla

Constructing distribution-free confidence intervals for the median, a classic problem in statistics has seen numerous solutions in the literature. While coverage validity has received ample attention, less has been explored about interval width. New ground is broken by investigating the width of these intervals under non-standard assumptions. Surprisingly, it is found that when properly scaled, the interval width converges to a non-degenerate random variable, unlike traditional intervals. The findings are also adapted to construct improved confidence intervals for general parameters, enhancing the existing HulC procedure. These advances provide practitioners with more robust tools for data analysis, reducing the need for strict distributional assumptions.

C1503: **Inference for projection parameters in linear regression: beyond $d = o(n^{1/2})$**
Presenter: **Woonyoung Chang**, Carnegi Mellon University, United States

Inference for the projection parameters is studied in the random-design linear regression model with increasing dimensions and under minimal distributional assumptions. This problem has been studied under a variety of assumptions in the literature. When the dimension d of the covariates is of smaller order than $n^{1/2}$, with n denoting the sample size, it is known that the traditional Wald confidence intervals based on the asymptotic normality of the least squares estimator and the sandwich variance estimator are asymptotically valid. A bias correction is developed for the least squares estimator, and the asymptotic normality of the resulting debiased estimator is proved as long as $d = o(n^{2/3})$, with an explicit bound on the rate of convergence to normality. Recent methods of statistical inference that do not require an estimator of the variance to perform asymptotically valid statistical inference are leveraged. It is discussed how the techniques can be generalized to increase the allowable range of d even further.

C1509: **Dimension-agnostic inference for M-estimation**
Presenter: **Kenta Takatsu**, Carnegie Mellon University, United States

Many statistical applications can be framed as the solution of stochastic optimizations whose objective function is estimated from data. This framework, widely known as M-estimation, includes maximum likelihood estimation, functional estimation, regression, and classification. Classical inferential methodologies for M-estimation often rely on strong assumptions about the underlying distribution as well as the asymptotic properties of the estimator. As a result, there is a severe shortage of inferential tools for high-dimensional settings. A simple method is proposed for constructing a confidence set for M-estimation, which remains valid without assumptions about the dimension of data. The proposed confidence set is based on a sample-splitting procedure, and conditions are provided under which its diameter converges at an optimal rate. Furthermore, the proposed method is applicable to non-standard problems where the inference has been notoriously difficult.

C1514: **Unified framework for inference using confidence sets for the CDF**
Presenter: **Siddhaarth Sarkar**, Carnegie Mellon University, United States
Co-authors: Arun Kuchibhotla

Traditional statistical inference methods often face limitations due to their reliance on strict assumptions. Moreover, these methods are typically tailored to specific assumptions, restricting their adaptability to any alternative set of assumptions. A unified framework is presented for deriving confidence intervals for various functionals (e.g., mean or median) under a broad class of user-specified assumptions (e.g., finite variance or tail behavior). Leveraging confidence sets for cumulative distribution functions (CDFs), this framework offers a principled and flexible inference strategy, reducing dependence on stringent assumptions and providing applicability in diverse contexts.

CO407 **Room BH (S) 2.02** **CLIMATE RISK UNCERTAINTY** **Chair: Maria Elena Bontempi**

C0200: **An econometric analysis of agricultural production, groundwater depletion and glacier thicknesses**
Presenter: **Alok Bhargava**, University of Maryland, United States

A framework is developed for assessing the sustainable production of staple crops in developing countries, taking into account their transpiration efficiency and groundwater depletion, which are critical in the wake of climate change. Moreover, recharging groundwater is important, and it is essential to monitor glacier thicknesses reflecting water storage, tackling the stochastic properties of glacier height changes assessed from remote sensing signals. First, the analysis of data on the production of rice, wheat, maize, sorghum, and pearl millet for 310 districts in India during 2000-2016, using dynamic random effects models, showed that greater rice output was significantly associated with higher well depths. Second, milk production was positively associated with well depths, indicating the need for higher water requirements for improving diet quality in developing countries. Third, an analysis of remote sensing data on changes in glacier heights for 0.5o x 0.5o grids in the Northern Hemisphere indicated large and significant declines over time; signal processing techniques induced trends in the means and variances of the series and glacier thickness reductions were generally small using in situ data. Overall, the results underscored the need for evidence-based policies that encourage the cultivation of water-efficient crops, monitoring of glaciers, and recharging groundwater to enhance food security in developing countries.

C0321: **Unobserved component models, approximate filters and dynamic adaptive mixture models**
Presenter: **Alessandra Luati**, Imperial College London, United Kingdom
Co-authors: Leopoldo Catania, Enzo DInnocenzo

State estimation in unobserved component models with parameter uncertainty is traditionally performed through approximate filters, where Gaussian distributions with given moments are employed to replace otherwise intractable conditional densities. The purpose is to re-examine signal-plus-noise models where parameter uncertainty is induced by a latent variable that may assume a fixed number of states. First, it is shown that, for these models, the approximate filters commonly adopted in the literature can be obtained as linear combinations of minimum variance linear unbiased estimators. Second, it is observed that they coincide with filters implied by a novel class of dynamic adaptive mixture models, where the parameters of a mixture of distributions evolve over time following a recursion that is based on the score of the one-step-ahead predictive distribution. Focusing on a robust specification, where the mixture components are Student's t distributions, it proves the existence, stationarity and ergodicity of the data-generating process as well as the invertibility of the filter and consistency and asymptotic normality of the maximum likelihood estimator of the static parameters. An application to a climate time series dataset is discussed, where the novel specification is compared with and shown to outperform robust score-driven filters and the related class of mixture autoregressive models.

C0669: **Spillover effect of private equity investment: Evidence from Italy**
Presenter: **Marzieh Abolhassani**, Free University of Bozen-Bolzano, Netherlands
Co-authors: Jan Ditzen

Private equity (PE) investments have emerged as a significant force in the financial landscape and global economy, influencing a wide array of industries and economic activities. Yet, the understanding of their economic implications is still limited. Spatial econometrics methods are applied to understand private equity spillover effects using novel panel data of 1,374 Italian PE-backed and matched control firms. Network models based on geography, industry and supply chain links are employed, as these networks can be important factors in the impact of private equity investment on target firm performance due to potential spillover. The hypothesis that a PE buyout in an industry will intensify the degree of competition in that industry with a negative effect on the performance of peers is tested. Additionally, PE buyouts are hypothesized to lead to a significant decline in productivity in the regions where the affected firms are located, due to localized disruptions in labor markets and supply chains. Results suggest, in essence, the significant negative effect of a PE buyout on firm productivity, indicating that the productivity decline has both regional and industry-wide repercussions, affecting a network of interconnected firms and regions beyond the initially impacted company.

C0743: **A new index of climate concern and identification of shocks: A proxy-SVAR approach**
Presenter: **Maria Elena Bontempi**, University of Bologna, Italy
Co-authors: Giovanni Angelini, Paolo Neri, Luca De Angelis, Marco Maria Sorge

While scientific consensus underlines the urgent need to address climate change, public perception plays a crucial role in driving policy and behavioural changes. Indeed, what is meant by climate change in common sense? Understanding public sentiment helps tailor communication strategies and make policies more effective. A new climate concern index (CCI) is presented, designed to gauge the heterogeneous levels of awareness of climate change among the population. Volumes of web searches are used for disaggregated queries that may produce worries in people. To understand variations in climate change perception, the analysis is based on U.S. and Italy, and different Italian regions, for the 2004-2024 period. Specific queries in the CCI capture the post-cognitive interpretation according to which, in order for an individual to develop negative affectivity toward climate change, that individual must first cognitively attribute personal experience with extreme weather to climate change. To measure the impact of climate concern shocks on macroeconomic outcomes, structural vector autoregressive (SVAR) models are used, which provide parsimonious characterizations of shock transmission mechanisms and track dynamic causal effects. As the identification of SVARs requires parameter restrictions on the matrix that maps the VAR disturbances to structural shocks, which are often implausible, a proxy-SVAR approach is used.

CO177 Room BH (S) 2.03 UNCERTAINTY AND INFORMATION FLOW IN FORECASTING AND FINANCIAL MARKETS Chair: Robert Kunst

C0745: **On the influence of the choice of seasonal adjustment method on forecasting national accounts aggregates across the EU**
Presenter: **Robert Kunst**, Institute for Advanced Studies, Austria
Co-authors: Martin Ertl, Adrian Wende

With quarterly national accounts aggregates, policy evaluation and forecasting often focus on seasonally adjusted time series. For the most part, two methods of seasonal adjustment are used today: Variants of the moving-average X-11 method with pre-specified filters and, alternatively, the SEATS method that is based on tentatively fitted ARIMA models. It is studied which of the two methods yields more accurate forecasts of annual targets after temporal re-aggregation, against two benchmark procedures: first, forecasting exclusively based on annual data; second, modeling quarterly year-on-year growth rates. These issues are investigated both empirically and with Monte Carlo simulations. For the simulations, data-driven time-series models are considered, both univariate and multivariate generating processes, ARIMA and VAR models among others, and various versions of seasonality generators. For the empirical investigation, the focus is on GDP and on related accounts aggregates in all EU economies.

C0854: **Commodity price uncertainty and macroeconomic dynamics**
Presenter: **Ines Fortin**, Institute for Advanced Studies, Austria
Co-authors: Jaroslava Hlouskova

Following the existing methodology, a new index is constructed, measuring commodity price uncertainty. The effects of commodity price uncertainty shocks are then examined on economic activity in the euro area, considering impulse response functions in a small VAR setup. Regime-dependent dynamics are allowed for, conditional on high or low commodity price uncertainty, and other uncertainty measures or commodity prices are also used to define regimes. In addition, the relative importance of commodity price uncertainty is assessed as a leading indicator for real and monetary developments in a forecasting analysis. The analysis focuses on certain commodity groups, in particular, selected industrial metals traded at the London Metal Exchange.

C0209: **Foreign economic policy uncertainty and the U.S. equity returns**
Presenter: **Mohammad Jahan-Parvar**, Federal Reserve Baord of Governors, United States
Co-authors: Jamil Rahman, Yuriy Kitsul, Beth Anne Wilson

The purpose is to document the predictive ability and economic significance of foreign economic policy uncertainty (EPU) for U.S. equity returns. After orthogonalizing global economic policy uncertainty (foreign EPU) with respect to the U.S. EPU, it is found that it has significant predictive power for aggregate stock returns and returns of portfolios constructed on size, investment, capital expenditure, and foreign sales in 6 to 12-months ahead horizons. It is found that foreign EPU shocks operate through firms' cash flow channels, they do not affect discount rates or equity premia, and that foreign EPU-sensitive firms respond to an adverse foreign EPU shock by altering their credit demand and investment outlays. However, their labor demand does not change much.

C0229: **The information advantage of banks: Evidence from their private credit assessments**
Presenter: **Mehdi Beyhaghi**, Federal Reserve Board, United States

In classic theories of financial intermediation, banks mitigate information frictions by monitoring and producing information about borrowers. However, it is difficult to test these theories without access to banks' private information. Supervisory data containing banks' private assessments of their loans' expected losses is used. It is shown that changes in expected losses predict firms' future stock returns, bond returns, and earnings surprises, and that banks use this information to allocate credit. Findings show that banks' information production and monitoring create an information advantage over financial markets, even among publicly traded firms.

CO199 Room BH (S) 2.05 FLEXIBLE ESTIMATIONS Chair: Jing Zhou

C0829: **Factor augmented tensor-on-tensor neural networks**
Presenter: **Xiufan Yu**, University of Notre Dame, United States

The prediction task of tensor-on-tensor regression in which both covariates and responses are multi-dimensional arrays (a.k.a., tensors) are studied across time with arbitrary tensor order and data dimension. Existing methods either focused on linear models without accounting for possibly nonlinear relationships between covariates and responses or directly employed black-box deep learning algorithms that failed to utilize the inherent tensor structure. A factor-augmented tensor-on-tensor neural network (FATTNN) is proposed to integrate tensor factor models into deep neural networks. It begins with summarizing and extracting useful predictive information (represented by the "factor tensor") from the complex structured tensor covariates. Then, it proceeds with the prediction task using the estimated factor tensor as input for a temporal convolutional neural network. The proposed methods effectively handle nonlinearity between complex data structures and improve over traditional statistical models and conventional deep learning approaches in both prediction accuracy and computational cost. By leveraging tensor factor models, our proposed methods exploit the underlying latent factor structure to enhance the prediction and drastically reduce the data dimensionality that speeds up the computation. Numerical results show the proposed methods achieve substantial increases in prediction accuracy and significant reductions in computational time compared to benchmark methods.

C1191: **Learning a directed acyclic graph with heteroscedastic errors**
Presenter: **Chunlin Li**, Iowa State University, United States

The purpose is to introduce a causal discovery method to learn the causal relations in a directed acyclic graph (DAG) with heteroscedastic errors. First, the model identifiability is derived for DAGs with heteroskedastic errors. Then, a new DAG reconstruction method called residual quantile (ResQ) estimation is proposed, which iteratively reconstructs the causal order of the variables. It is proven that ResQ enjoys desirable statistical properties such as reconstruction consistency in both low- and high-dimensional cases and estimation robustness to heteroskedastic/heavy-tailed/contaminated errors. The theoretical properties have been substantiated by the synthetic experiments and applications to real-world causal benchmark datasets, where ResQ compares favorably against state-of-the-art competitors.

C1229: **High-dimensional regression: Model averaging and inference**
Presenter: **Lise Leonard**, UCLouvain, Belgium
Co-authors: Eugen Pircalabelu, Rainer von Sachs

With the advent of technology and the proliferation of data collection methods, researchers now have access to vast amounts of data. High-dimensional regression models are designed to handle datasets with more predictors than observations. However, these methods, such as the Lasso, depend on unknown tuning parameters. A procedure for high-dimensional model averaging is proposed that allows inference, with the goal of eliminating the difficult choice of the tuning parameter and obtaining an asymptotically Gaussian estimator. The main feature of the procedure is to pool together information from multiple estimators to obtain a single, final estimator. A strategy based on the debiased Lasso estimator is proposed to aggregate regression coefficients to reduce the prediction risk of the estimation and to eliminate the influence of the tuning parameter. Theoretical results on the distribution and the prediction risk of the method are presented. In particular, the asymptotic normality of the estimator is shown even after the aggregation and the optimality of the prediction loss. The performance of the method is illustrated through numerical simulations and an application on a real, high-dimensional dataset.

C1515: **Dependence of drought characteristics: Parametric and non-parametric copula approach**
Presenter: **Ozan Evkaya**, Edinburgh University, United Kingdom

To better understand and monitor the effects of drought, various methods have been developed in recent decades to quantify drought characteristics. Understanding drought characteristics is essential for conducting an in-depth examination of its impacts on a specific area. That specifically requires examining the specific characteristics of drought, such as its duration or severity and including the association between these characteristics. In that respect, it is crucial to model the joint behavior of these drought characteristics. The endeavor is to investigate univariate and bivariate drought indices using both parametric and nonparametric copula techniques. For that purpose, drought characteristics, such as duration, severity, mean intensity, and peak intensity, are analysed based on different drought indices. The dependence among the main characteristics is evaluated, and corresponding bivariate return period calculations are investigated. The used data set is retrieved from monthly meteorological observations collected at five different stations in Konya, located in the Central Anatolia Region of Turkey. As explored, parametric or nonparametric copula usage may differ slightly based on the extreme drought cases. The findings of the study examine the suitability of both parametric and nonparametric dependence settings for a specific region by testing across different weather stations.

CO092 **Room BH (SE) 2.05** STATISTICAL INFERENCE, MODELING, AND OPTIMAL DESIGN **Chair: Subir Ghosh**

C1215: **Online Bayesian model averaging for streaming data**
Presenter: **Joyee Ghosh**, The University of Iowa, United States
Co-authors: Aixin Tan, Lan Luo

There is an increasing prevalence of streaming data generation in diverse fields like healthcare, finance, social media, and weather forecasting. In order to acquire helpful insights from these massive datasets, timely analysis is essential. The streaming data is assumed to be analyzed in batches. Traditional offline methods, which involve storing and analyzing all individual records, can be repeatedly applied to the cumulative data but encounter significant challenges in storage and computing costs. Existing online methods offer faster approximations, but most methods neglect model uncertainty, causing overconfidence and instability. To bridge this gap, novel online Bayesian approaches are proposed that incorporate model uncertainty within a Bayesian model averaging (BMA) framework, for generalized linear models (GLMs). Computationally efficient methods are proposed to update the posterior, with individual records from the latest batch of data and summary statistics from previous batches. Simulation studies and real data demonstrate that the methods can offer much faster analysis compared to traditional methods, with no substantial drop in accuracy.

C1252: **Responsibly emboldening predictions via boldness-recalibration**
Presenter: **Adeline Guthrie**, Virginia Tech, United States
Co-authors: Christopher Franck

Probability predictions are essential for informing decision-making across many fields. The purpose is to enable better decision-making by improving probability forecasts in terms of their calibration and boldness - which are essential properties of useful forecasts. A Bayesian model selection approach is proposed to assess calibration and a strategy for boldness-recalibration that enables practitioners to responsibly embolden predictions subject to their required level of calibration. Specifically, the user is allowed to pre-specify their desired posterior probability of calibration and then maximally embolden predictions subject to this constraint. The key of these methods is demonstrated via real-world case studies pertaining to the prediction of housing foreclosures and ice hockey games.

C1212: **Optimal design construction: A comparative study of various methods**
Presenter: **Akram Mahmoudi**, Jonkoping International Business School, Sweden
Co-authors: Saumen Mandal

Owing to the widespread application of optimal designs, the motivation is to compare different methods of optimum designs. In this regard, some optimization algorithms are used to construct approximate optimal designs. Some of these methods are gradient-based, while others are gradient-free algorithms. The studied methods include a class of multiplicative algorithms, simulated annealing and optimization with inequality constraints-Barrier methods. Optimization algorithms are simulated in three models: a quadratic model, a practical model in chemistry, and a model with two design variables. The simulation study is done under different scenarios to explore the performance of the algorithms. Then, the strengths and weaknesses of the methods are explored, and a comparison between methods is done.

C1385: **Confidence regions when the parameter is near the boundary**

Presenter: **Karl Oskar Ekvall**, University of Florida, United States

The focus is on recent advances in the theory and methods for constructing reliable confidence regions when the parameter may be near the boundary of the parameter set. For reasons to be discussed, constructing such confidence regions is often substantially more difficult than the testing of boundary points. Recent work shows a connection between boundary problems and a singular Fisher information that can sometimes be leveraged to provide reliable conference regions. However, it is unclear how far that approach generalizes, and other arguments appear more useful in some settings. Examples using variance components and mixed models illustrate the general theory and also suggest several directions for future research. For example, the need for reliable confidence intervals for variance components suggests studying settings where both a parameter of interest and nuisance parameters can be near the boundary.

CO020 Room BH (SE) 2.09 INDIVIDUAL HETEROGENEITY FOR MACROECONOMIC FLUCTUATIONS Chair: Michele Lenza

C1147: **Measuring the effects of aggregate shocks on unit-level outcomes and their distribution**

Presenter: **Stephanie Ettmeier**, University of Bonn, Germany

Co-authors: Frank Schorfheide

The effect of aggregate shocks is studied on micro-level outcomes. A cross-sectional units vector autoregression (csuVAR) that combines aggregate variables with unit-level outcomes and earnings in the application is developed and estimated. The csuVAR also allows reconstruction of the cross-sectional distribution from the unit-level outcomes. The csuVAR is contrasted with a functional VAR model (fVAR) that is designed to directly track the evolution of macroeconomic aggregates and a cross-sectional distribution, but not individual units. In an empirical application, we examine the effect of productivity shocks on the unit-level labor earnings dynamics in Germany, using a panel data set constructed from the Sample of Integrated Labor Market Biographies (SIAB) published by the Institute for Employment Research (IAB) of the German Federal Employment Agency.

C1186: **Long-run neutrality meets reality: Innovation and monetary policy**

Presenter: **Michaela Elfsbacka Schmoller**, European Central Bank, Germany

The focus is on studying the effect of monetary policy shifts on innovation using a large, representative survey of German firms with unique, granular information. The findings show that interest rate hikes in 2022-2023 have significantly affected innovation investment, with about 30% of firms reducing expenditures, of which about 40% cut to zero. Rich information is further exploited from hypothetical, incremental rate change scenarios to analyze potential asymmetries and non-linearities in the transmission mechanism of monetary policy to innovation. The effect of forward guidance is studied, which suggests that announcements influence firms' innovation decisions and longer-term aggregate supply. The results challenge the long-run neutrality of monetary policy, a key assumption in New Keynesian models, and suggest that monetary policy significantly affects firms' innovation investment and, thus, technology growth and aggregate supply over the medium term and beyond.

C1144: **On the need of firm data to understand macroeconomic dynamics**

Presenter: **Michele Lenza**, European Central Bank, Germany

Co-authors: Ettore Savoia

The role of heterogeneity in the revenues of individual firms for euro area macroeconomic dynamics is studied. To this end, two models are specified: a standard aggregate vector autoregressive model (VAR) and a heterogeneous VAR (HVAR). The VAR model includes only aggregate data, while the HVAR model also incorporates the feedback loop between firms' revenue distribution and aggregate variables. The results demonstrate that the behavior of the firms' revenue distribution plays a significant role in explaining the dynamics of key euro area macroeconomic variables.

C1609: **Firms heterogeneity and aggregate fluctuations: What can we learn from machine learning**

Presenter: **Luigi Pollio**, UMBC, United States

Co-authors: Simone Pesce, Marco Errico

The heterogeneous sensitivity of firms to aggregate fluctuations affects business cycle dynamics. The aim is to examine how firms' outcomes (sales, debt, investment, market value) respond to aggregate fluctuations (business cycle, monetary policy, uncertainty, oil shocks) based on eight firm characteristics using the generalized random forest algorithm. Analyzing Compustat micro-level data, three key findings are documented: (1) while linear OLS captures the average effect, there is substantial cross-sectional heterogeneity in firm sensitivities; (2) the importance of firm characteristics varies across outcomes and shocks, with non-financial characteristics explaining more of the heterogeneity; (3) firm sensitivity to aggregate fluctuations shows non-linear patterns based on characteristics. An aggregation theory is developed, and the estimated model is used to generate counterfactual firm-level sensitivities. The findings reveal that (1) heterogeneity in firm sensitivity amplifies aggregate responses to macroeconomic variables; (2) non-linearity at the micro-level has little effect on aggregate responses; (3) non-financial characteristics drive aggregate responses more than financial ones; and (4) state-dependent heterogeneity plays a minor role in influencing aggregate responses.

CO354 Room BH (SE) 2.10 SPATIAL STATISTICS AND ITS APPLICATIONS Chair: Rajarshi Guhaniyogi

C1651: **Analyzing structural breaks in the spatial network of real estate dynamics: A study of UK property transactions**

Presenter: **Soudeep Deb**, Indian Institute of Management Bangalore, India

Co-authors: Archi Roy, Satyaki Basu Sarbadhikary

The real estate market is a complex system affected by various economic, social, and environmental factors. Detecting structural changes is crucial for understanding market dynamics, adapting to trends, mitigating risks, and making informed investment and policy decisions. The aim is to analyze temporal changes in the UK real estate market using weekly data at the MSOA (Middle Layer Super Output Areas) level. A two-stage methodology is applied: first, using LISA (Local Indicators of Spatial Association), significant clusters of high or low real estate prices and spatial outliers are identified. These clusters are then integrated into a network structure, incorporating geographical distance as an attribute. In the second stage, network Laplacians are used to detect structural breaks over time. By examining the singular values of the networks at different time points, vectors are identified, indicating smooth or abrupt changes. This approach uncovers shifts in UK house prices, offering valuable insights into regional trends, demand changes, and economic impacts, benefiting urban planners and investors in their strategies.

C1569: **Matrix-free conditional simulation of Gaussian random fields**

Presenter: **Somak Dutta**, Iowa State University, United States

Co-authors: Debashis Mondal

Conditionally simulating Gaussian random fields given a large set of observations presents significant computational challenges. These challenges stem from the necessity to compute and store large matrices or their explicit factorizations. Typically, these requirements grow at a super-linear rate with the number of observations. A novel approach that relies on matrix-free rectangular roots of precision matrices is introduced. This approach's applicability is demonstrated in a widely used class of spatial models and outperforms existing methods that use sparse matrix factorizations or other techniques. The effectiveness of the method is illustrated by applying it to analyze groundwater arsenic contamination in Bangladesh and ocean temperature measurements from Argo floats.

C1598: **Statistical analysis of extreme geostatistical data: Challenges and advances**

Presenter: **Dora Prata Gomes**, NOVA.ID.FCT FCT-UNL, Portugal

Co-authors: Clara Cordeiro, Manuela Neves

The quantification and characterization of extreme meteorological events are crucial due to their significant impacts on human life, agricultural productivity, and economic stability. Extreme value theory (EVT) is a theory for modeling and quantifying events that have a very low probability of occurrence but can result in significant impacts. Given that a single extreme weather event can influence multiple locations, the assumption of temporal independence is often unrealistic. While stationary and weakly dependent series share the same limiting distribution as independent ones, their parameters are affected by dependence. Therefore, when analyzing data from multiple locations, it is essential to account for spatial dependence appropriately. Traditional geostatistics, which predominantly relies on Gaussian distribution, is inadequate for modeling tail behavior. The application of max-stable processes is explored, a natural extension of multivariate extremes in a spatial context. An exploratory analysis of annual maximum monthly precipitation data recorded from 1941 to 2023 across 19 locations in northern Portugal is conducted. The results include reviewing dependence measures, estimating parameters of interest, and simulating rainfall prediction maps. The R software will be explored, and packages and functions related to this topic will be used.

C0197: **Bayesian modelling for spatially misaligned health areal data**
Presenter: **Silvia Liverani**, Queen Mary University of London, United Kingdom

The objective of disease mapping is to model data aggregated at the areal level. In some contexts, however (e.g. residential histories, general practitioner catchment areas), when data arising from a variety of sources, not necessarily at the same spatial scale, it is possible to specify spatial random effects or covariate effects at the areal level, by using a multiple membership principle. The purpose is to investigate the theoretical underpinnings of this application of the multiple membership principle to the CAR prior, in particular with regard to parameterization, properness and identifiability, and the results of an application of the multiple membership model to diabetes prevalence data in South London are presented, together with strategic implications for public health considerations.

CO309 Room BH (SE) 2.12 STATISTICAL METHODS FOR ANALYZING HIGH-DIMENSIONAL CANCER DATASETS Chair: Subharup Guha

C0369: **Accounting for network noise in graph-guided Bayesian modeling of high-dimensional-omics data**
Presenter: **Wenrui Li**, University of Connecticut, United States
Co-authors: Changgee Chang, Suprateek Kundu, Qi Long

There is a growing body of literature on knowledge-guided statistical learning methods for the analysis of structured high-dimensional data (such as genomic and transcriptomic data) that can incorporate knowledge of underlying networks derived from functional genomics and functional proteomics. These methods have been shown to improve variable selection and prediction accuracy and yield more interpretable results. However, these methods typically use graphs extracted from existing databases or rely on subject matter expertise, which are known to be incomplete and may contain false edges. To address this gap, a graph-guided Bayesian modeling framework is proposed to account for network noise in regression models involving structured high-dimensional predictors. Specifically, two sources of network information are used, including the noisy graph extracted from existing databases and the estimated graph from observed predictors in the dataset at hand, to inform the model for the true underlying network via a latent scale modeling framework. This model is coupled with the Bayesian regression model with structured high-dimensional predictors involving an adaptive structured shrinkage prior. An efficient Markov chain Monte Carlo algorithm is developed for posterior sampling. The advantages of the method over existing methods are demonstrated in simulations and through analyses of a genomics dataset and another proteomics dataset for Alzheimer's disease.

C0375: **Harnessing sociocultural similarities between diverse populations to identify determinants of cancer screening use**
Presenter: **Jaya Satagopan**, Rutgers School of Public Health, United States
Co-authors: Shromona Sarkar

Screening is effective in detecting cancer early when it is easier to treat and the chances of survival are better. However, there is considerable racial disparity in the use of cancer screening. Standard approaches to identify factors associated with these disparities begin by stratifying the data on race and then fitting regression models. However, strata based on discrete labelling of race, such as non-Hispanic White, non-Hispanic Black and so on, are likely to be oversimplified when attempting to interpret screening use in diverse populations. Further, the sample sizes of under-studied populations such as non-Hispanic Black, Hispanic, and Asian subgroups are often small, which diminishes the power to detect factors that determine screening use in some, but not necessarily all, strata. To address these challenges, a latent class approach is developed and implemented to stratify individuals according to their socio-cultural similarities, regardless of their race/ethnicity. The properties of the approach are illustrated, and its benefits are demonstrated by conducting comparisons with standard methods using multiple years of data from the National Health Interview Survey to examine disparities in mammogram screening use in diverse populations.

C0378: **A pseudo-value approach to causal deep learning of semi-competing risks**
Presenter: **Stephen Salerno**, Fred Hutchinson Cancer Center, United States
Co-authors: Yi Li

While mortality is often the main focus of cancer studies, non-fatal events (i.e., disease progression) can vitally impact patient outcomes. Recurrence after curative treatment is a crucial endpoint in lung cancer, affecting second-line treatment options. Estimating the de-confounded effect of an intervention on disease recurrence is a key aspect of assessing cancer treatments. However, semi-competing risks complicate causal inference when death prevents disease recurrence. Existing approaches for estimating causal quantities for semi-competing risks rely on complex objective functions with often strong assumptions. To address these challenges, a deep learning approach is proposed for estimating the causal effect of treatment on non-fatal outcomes in the presence of dependent censoring. The three-stage approach involves estimating the non-fatal survival function, constructing jackknife pseudo-survival probabilities at fixed time points, and fitting a deep neural network to estimate the effect of treatment. The pseudo-survival probabilities serve as target values for developing causal estimators that are consistent and do not rely on assumptions like proportional hazards, which enables estimating survival average causal effects through direct standardization. The approach is evaluated through numerical studies, and it is applied to the Boston Lung Cancer Study to estimate the effect of surgical tumor resection in patients with early-stage non-small cell lung cancer.

C0875: **Identifying genes associated with disease outcomes using joint sparse canonical correlation analysis**
Presenter: **Diptavo Dutta**, National Cancer Institute, United States

Genomic and epigenomic changes can have pivotal effects on cancers, and joint analyses of such multimodal data can identify novel biomarkers for cancer-related outcomes. Joint sparse canonical correlation analysis (jsCCA) is proposed to identify an ensemble of copy number aberrations (CNAs), methylation sites and gene expressions relevant to tumor outcomes. JsCCA detects orthogonal gene modules correlated with sets of methylation sites, which in turn are correlated with sets of CNA. Analysis of data on 515 kidney cancer patients from the TCGA-KIRC found eight gene modules associated with methylation sites and groups of proximally located CNA sites. ASAH1 gene is identified, trans-regulated by CNA and methylation sites, to be associated with tumor stage. Quantifying the overall effect of gene modules revealed that two gene components have significant interaction with smoking and represent distinct biological functions, including inflammatory responses and hypoxia-regulated pathways. The results indicate that methods like jsCCA are warranted for integrative analysis of multimodal data in cancer genomics to identify interpretable, novel, and clinically relevant molecular targets.

CC479 Room S-1.01 SOFTWARE **Chair: Masayuki Hirukawa**

C1595: **Safety first: Design-informed inference for treatment effects via the propertee package for R**
Presenter: **Ben Hansen**, University of Michigan, United States
Co-authors: Joshua Errickson, Joshua Wasserman, Adam Sales

When treatments are allocated by cluster, it is vital for correct inference that the clustering structure be tracked and appropriately attended to. In randomized trials and observational studies modeled on RCTs, clustering is determined at the early stage of study design, with subtle but important implications for the later stage of treatment effect estimation. A first contribution of our "propertee" R package is to make analysis safer by providing self-standing functions to record treatment allocations, with the thus-encoded study design informing subsequent calculations of inverse probability weights, if requested, and of standard errors. A second contribution is to facilitate the use of precision-enhancing predictions from models fitted to external or partly external samples. The user experience is kept simple by adapting familiar R mechanisms such as predict(), lm(), offset(), the sandwich package and summary.lm(); it uses stacked estimating equations under the hood. The propertee package makes it easy and safe to produce Hajek- or block fixed effect estimates with appropriate standard errors, even in the presence of grouped assignment to treatment, repeated measures, subgroup-level estimation and/or covariance adjustment.

C1587: **Analyze of count longitudinal data with random effects using R packages, cold, lme4 and glmmML**
Presenter: **Maria Helena Goncalves**, FCiencias.ID, Associacao para a Investigacao e Desenvolvimento de Ciencias (Portugal), Portugal
Co-authors: Maria Salome Cabral

Longitudinal count data are commonly encountered in both experimental and observational studies across all disciplines. In these studies, repeated measurements are made on the same subject across occasions in one or more treatment groups, and correlation is usually present among response variables for a given subject. The generalized linear mixed models (GLMMs) account for that correlation by the inclusion of random effects in the linear predictor. However, in GLMMs it is assumed that the observations of the same subject are independent conditional to the random effects and covariates, which may be not true. For fitting GLMMs, R has available the packages lme4 and glmmML, at the least. The methodology implemented in cold R package inference is based on the likelihood approach, serial dependence is assumed to be of Markovian type, and it is considered as the basic stochastic mechanism. The serial dependence AR1 incorporated in the random effects model in cold allows dependence between repeated measures in terms of numerical analysis, which is ignored in the traditional approach (GLMM) implemented in the lme4 and glmmML. The R packages lme4 and glmmML only allow an independent structure, and glmmML only allows a random effect in the intercept. A real dataset is used to compare the aforementioned R packages.

C1677: **Functional data clustering in R**
Presenter: **Manuel Oviedo de la Fuente**, University of a Coruña, Spain
Co-authors: Manuel Febrero-Bande

Functional data clustering aims to identify heterogeneous patterns within continuous functions such as curves, images and surfaces. The remarkable growth in the application of functional data clustering highlights the need for a systematic approach to developing efficient clustering methods and scalable and user-friendly algorithms. We present the main functional data clustering methods available in R software, with a particular focus on the new version of the fda.usc library. Key functional clustering methods such as hierarchical clustering, DBSCAN, mean shift and k-means are highlighted. In addition, we illustrate methods for selecting the optimal number of clusters or evaluating cluster quality in functional data contexts, using both simulated and real data scenarios.R offers several notable packages for FDA (see CRAN Functional Data Task View), including fda, which serves as the foundation for many subsequent packages. One of these is fda.usc, which builds on some of the fda utilities while incorporating additional nonparametric techniques, among others. The focus is on the innovations introduced in the latest version, which provide a valuable resource for the scientific community by simplifying the analysis of complex functional data in an accessible and reproducible manner.

C1396: **"clustglm" and "clustord": R packages for clustering with covariates for binary, count, and ordinal data**
Presenter: **Louise McMillan**, Victoria University of Wellington, New Zealand
Co-authors: Daniel Fernandez, Shirley Pledger, Richard Arnold, Ivy Liu, Murray Efford

Two R packages are presented for model-based clustering with covariates. Both packages can perform clustering and biclustering (clustering observations and features simultaneously, for example). Both use likelihood-based methods for clustering, so users can compare models using AIC and BIC to assess relative goodness of fit. The models in both packages use linear predictor terms, so they look more like regression models than clustering models. This allows the inclusion of regression-style covariates alongside clustering effects. Both "clustglm" and "clustord" can include the effects of numerical or categorical covariates alongside cluster effects or can fit pattern-detection models that include individual-level effects alongside cluster effects. For example, when applied to presence/absence data, sites and species are clustered while also taking into account any single-species effects and any additional covariates."clustglm" is designed for binary and count data. It uses "glm" and can accommodate balanced and non-balanced designs. "clustord" is designed for ordinal categorical data. It can fit the proportional odds model or the ordered stereotype model, a more flexible model whose fitted parameters can reveal when two ordinal categories are effectively equivalent to each other. The use of "clustglm" and "clustord" is illustrated with ecological and survey datasets.

CC483 Room BH (SE) 1.05 EMPIRICAL FINANCE **Chair: Nicola Loperfido**

C1046: **Probability forecasts: A simple albeit powerful predictor for hedge fund returns**
Presenter: **Michail Karoglou**, Aston Business School, United Kingdom
Co-authors: Emmanouil Platanakis, Dimitrios Stafylas

The use of simple probability forecast risk measures (PFRMs) is proposed to capture forward-looking information for various negative and extreme events for hedge funds. It is shown that individual PFRMs and various aggregations using popular machine learning methods (PLS, sPCA, C-Lasso, and C-Enet) can predict the total hedge funds' return significantly out-of-sample and outperform popular predictors. This strong predictability power is maintained for many hedge fund categories and remains robust to several additional checks.

C1537: **An empirical comparison between investing strategies: Maximum diversification versus minimum risk**
Presenter: **Pierpaolo Uberti**, University of Milano-Bicocca, Italy
Co-authors: Maria-Laura Torrente

In well-defined experimental settings, the out-of-sample performance of two asset allocation paradigms is evaluated: minimum risk and maximum diversification. Specifically, for each given risk measure, the optimal minimum risk allocation is compared with the allocation obtained maximizing a portfolio diversification measure induced by the same risk measure. The experiment is performed in an out-of-sample long-only framework, considering proportional transaction costs and different lengths of the estimation window and the holding period. The strategies are compared in terms of numerical stability, return, Sharpe ratio and risk measured through the same risk measures used for the calculation of the optimal allocation: variance of returns, mean absolute deviation, Value-at-risk and expected shortfall both at a significance level of 1% and 5%. It is shown that the maximum diversification strategies are very competitive, if not better in general than the risk minimization allocations. This result confirms well-known empirical findings of naive investment strategies that are difficult to beat in practice.

C1635: **Volatility transmission between commodity option and futures markets**
Presenter: **Gabriel Power**, Laval University, Canada

Co-authors: Marie-Helene Gagnon, Constant Aka

Derivatives markets are essential to price discovery in finance. The bidirectional volatility relationship is examined between options and futures markets from 11/2011 to 08/2022 for economically important commodities such as crude oil, natural gas, gold, wheat, corn and lean hogs in the USA (Chicago Mercantile Exchange). Information diffusion and risk spillovers are studied between these assets using random forest models, impulse response functions, and spillover measures, including the approach in a prior study. First, it is found that futures realized volatility immediately affects option volatility. Second, option volatility affects futures volatility less quickly but with a much longer-lasting impact. These results confirm the bidirectional relationship. The spillover analysis shows predominant self-driven volatility across most commodities, with notable net spillovers from options to futures. Finally, predictive analysis using random forests reveals that options markets generally lead futures markets in terms of providing useful information for predicting volatility. They provide significantly more accurate futures volatility predictions and allow for superior economic gains based on simple but feasible trading strategies presented.

C1494: **Political competition, democracy and financial development: Cross-country evidence**
Presenter: **Alfonsina Iona**, Queen Mary University of London, United Kingdom
Co-authors: Leone Leonida, Dawit Zerihun Assefa

Despite extensive research in politics and finance, the impact of political competition on countries' financial development has remained underexplored. The hypotheses of prior studies are integrated into a single reduced-form model to examine the relationship between political competition and financial development. These hypotheses are tested using a sample of over 100 countries from 1980 to 2020, employing both annual data and five-year averages. Findings show that in OECD countries, the relationship between political competition and financial development follows an S-shaped curve, while the entire sample exhibits a U-shaped pattern. These results suggest that in democratic countries, mechanisms similar to those described in a prior study prevail, whereas in autocratic countries, a political replacement effect akin to that proposed in another study is at play. The conclusions are robust across various model specifications, econometric estimators, and measures of financial development and political competition.

CC430 Room BH (S) 2.01 TIME SERIES ECONOMETRICS Chair: Weining Wang

C0392: **Going upstream: Responses of drilling activity to global oil markets**
Presenter: **Wenqi LI**, Chongqing University, China
Co-authors: Bao Nguyen

Understanding supply responses in the oil market and guiding strategic decisions in the energy sector is crucial, with drilling activity playing a key role. To analyze the dynamic adjustments of drilling operations to market signals, particularly oil price shocks, experiments are conducted using an innovative methodological approach that integrates oil stocks both above and below ground. A structural vector autoregressive (SVAR) model, augmented by Bayesian methodology, is employed. The analysis reveals the pronounced sensitivity of drilling activity to these shocks, highlighting the critical role of subterranean oil stocks, which significantly influence market expectations and pricing structures. Counterfactual analysis of scenarios, such as the COVID-19 pandemic and significant geopolitical disruptions, explores potential alternative trajectories of drilling activity under varying economic conditions, pinpointing crucial periods of deviation from projected drilling patterns. Furthermore, marked differences in drilling responses pre- and post-shale oil revolution are indicated, emphasizing the transformative impact of technological advancements on drilling methodologies.

C1238: **New proposal for seasonal adjustment of long time series**
Presenter: **Cheyenne Amoroso**, University of Coruna, Spain

A common economic task is the seasonal adjustment of time series, which involves removing the seasonal component from the data. Currently, at the National Statistics Institute (INE), this task is performed using the Tramo-Seats methodology. The time series currently being processed extend over many years, which generally complicates the identification of a single reg-ARIMA model that adequately describes the behavior of the entire series. Moreover, the data suggest that the behavior of the series has changed following the 2008 crisis. Motivated by the aforementioned issues, new general methodologies are suggested to perform seasonal adjustment in a long time series with two identified models and a transition period. The series before and after the event are considered modellable using ARIMA models, while the transition period is modeled as a weighted average of the other two events through a time-dependent weighting function. The proposals are assessed through an exhaustive simulation study aimed not only at verifying the gains compared to classical methodologies but also at evaluating their robustness in unfavorable scenarios.

C1554: **Forecasting with dynamic factor models estimated by partial least squares**
Presenter: **Samuel Rauhala**, University of Turku, Finland

Dynamic factor models (DFMs) have found great success in nowcasting and short-term macroeconomic forecasting. The factor loadings are typically estimated cross-sectionally, for example, with principal components. This ignores whether or not the factors have predictive capabilities. Two alternative approaches are suggested, using partial least squares, which takes the time series structure better into account. The first one is close akin to a large vector autoregression, and the second is more akin to a conventional DFM. Monte Carlo exercises are conducted, and it is found that in finite samples, this method outperforms the typical methods, such as the two-step estimator and quasi-maximum-likelihood.

C1681: **Forecasting cryptocurrency returns with a sparse dynamic factor model**
Presenter: **Tatsuma Wada**, Keio University, Japan
Co-authors: Akihiko Noda

The predictability of cryptocurrencies is assessed using a sparse dynamic factor model (DFM). The motivation for using this model stems from the categorization of cryptocurrencies into two groups: stablecoins and non-stablecoins. Stablecoins are pegged to stable assets, such as fiat currencies, making them less susceptible to speculative transactions, while non-stablecoins are not, leading to more volatile prices. Given these distinctions, it is reasonable to consider several factors that influence the prices of various cryptocurrencies. Our findings indicate that the sparse DFM outperforms the random walk model. However, the comparison with the vector autoregressive (VAR) model is largely inconclusive. Notably, however, when additional financial variables are incorporated into the VAR, the sparse DFM demonstrates superior performance. The model also outperforms the VAR when the rolling window is relatively small and the forecasting horizon is moderate. These findings suggest that the sparse DFM is robust in small sample sizes and is particularly well-suited for forecasting cryptocurrency prices in the near to mid-term future.

CC495 Room BH (SE) 2.01 DATA ANALYSIS AND EMPIRICAL STUDIES Chair: Joshua Cape

C1419: **A comparison of numerical maximum likelihood and noise-contrastive estimation for unnormalized statistical models**
Presenter: **Marco Bee**, University of Trento, Italy
Co-authors: Flavio Santi

Unnormalized statistical models are commonly used in many fields of statistics. However, their maximum likelihood estimation is difficult because the intractable normalizing constant depends on the unknown parameters. One way of proceeding is based on the maximization of an approximation of the likelihood obtained by means of a numerical estimate of the normalizing constant, even though the estimation error caused by the numerical evaluation of the normalizing constant is difficult to quantify. A popular alternative is noise-contrastive estimation (NCE), where a classification algorithm tries to discriminate synthetic and observed data and the normalizing constant is estimated alongside the other parameters of the model. The two approaches outlined above are used in two examples: the estimation of the parameters of the Bingham distribution for directional data on the unit sphere and the estimation of a dynamic mixture for loss data. For NCE, the well-known issue of finding an "optimal" noise distribution is also studied. Numerical maximum likelihood is usually more efficient in terms of root-mean-squared-error, whereas noise-contrastive estimation is faster and less affected by convergence problems.

C1650: **Quantifying the intrinsic data quality of process data**

Presenter: **Chong Dae Kim**, TH Koeln (Technische Hochschule Koeln), Germany

In an increasingly digitized world, the role of data in all its forms is essential. This importance aligns with the growing emphasis on data sharing. To allow data consumers to assess the quality of data in advance, it should be recorded in a manner that is easily and clearly reproducible. Following an extensive literature review that did not yield suitable methods, a framework for quantifying the intrinsic data quality of process data is presented in the form of time series. For this purpose, the dimensions of intrinsic data quality presented are individually made mathematically assessable and implemented into a Python framework for testing, validating and producing results. The quantification method presented is validated using measurement data obtained from a gear test rig, accompanied by a systematic evaluation of the dataset. It is demonstrated that the methodology considers various dimensions in addition to accuracy. Nevertheless, score enhancement is achievable through data preprocessing. Further research should explore the application of the methodology across diverse datasets and quantify additional dimensions of data quality.

C1600: **Comparisons of variable selection methods in mixtures of linear regression models**

Presenter: **Susana Faria**, Universidade do Minho, Portugal

Co-authors: Ana Moreira

Finite mixture regression models provide a flexible tool for modelling data that arise from a heterogeneous population, where the relationship between the dependent variable and the explanatory variables varies across different subpopulations. In the applications of these models, a large number of explanatory variables is often considered; for this reason, variable selection assumes great relevance for mixture models. However, since all subset selection methods are computationally intensive, more efficient methodologies were developed to overcome this problem, such as methods based on penalty functions. The least absolute shrinkage and selection operator (LASSO) method, the Adaptive Least Absolute Shrinkage and Selection Operator (ALASSO) method and the relaxed least absolute shrinkage and selection operator (RLASSO) method are some examples of these methods. The problem of variable selection in mixtures of linear regression models in the presence of a large number of explanatory variables is analyzed by comparing the performance of LASSO, ALASSO and RLASSO in the selection of explanatory variables, employing the expectation-maximization and classification expectation-maximization algorithms for parameter estimation. An extensive simulation study reveals how different scenarios affect the performance of these algorithms and penalization functions in selecting explanatory variables.

C1560: **Odds ratio for assessing the risk of crime associated with darkness**

Presenter: **Ezgi Erturk**, University College London, United Kingdom

Co-authors: Jemima Unwin Teji, Peter Raynham

The aim is to explore whether manipulating the control period does not affect the results in estimating the crime risk associated with darkness. The Metropolitan police services dataset for London between 2013 and 2019 was used. Comparing crime rates that took place in light in one week and dark in the other week at the same clock time (spring and autumn clock change) or vice versa is defined as the "Test period." Different control periods were created to compare the changes in the crime counts in the test period and to minimise other factors that may encourage crime. Therefore, 45-minute, 1-hour, and 2-hour lengths of control periods on the odd ratio calculation were explored. Although, in the 45-minute duration, the odds ratio of criminal damage, robbery, and motor vehicle offences increased by 7%, 11%, and 10% more compared to the one-hour period, the significance tests showed that all p-values were well above the conventional threshold of 0.05. The only noticeable change was in the odds ratio for robbery from the person between the 45-minute and 2-hour control periods, with a p-value of 0.089. However, this result was still not statistically significant. It shows that even through the control period, there are no significant statistical changes. All variables demonstrate that darkness has an impact on crime. It is important to understand the crimes and areas in which electric lighting design is likely to have an effect.

Authors Index