PROGRAMME AND ABSTRACTS

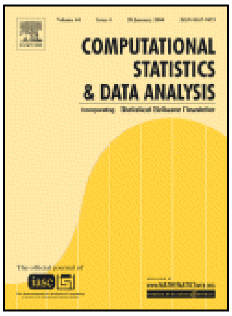
6th International Conference on

Econometrics and Statistics (EcoSta 2023)

http://cmstatistics.org/EcoSta2023

Waseda University, Tokyo, Japan

1-3 August 2023



ISBN: 978-9925-7812-2-5

Co-chairs:

Cathy W. S. Chen, Donggyu Kim, Michele Guindani and Yan Liu.

EcoSta Editors:

Ana Colubi, Erricos J. Kontoghiorghes and Manfred Deistler.

Scientific Programme Committee:

Soutir Bandyopadhyay, Joshua Cape, Alessia Caponera, Keith Chan, Ming-Chung Chang, Kun Chen, Hyunkeun Cho, Boris Choy, Yeonseung Chung, Takeshi Emura, Tsai-Hung Fan, Lisa Goldberg, Il Do Ha, Tadao Hoshino, Hsin-Cheng Huang, Binyan Jiang, Feiyu Jiang, Moritz Jirak, Luke Keele, Esra Kurum, Kuang-Yao Lee, Sangyeol Lee, Cheng Li, Daoji Li, Gen Li, Tsung-I Lin, Zudi Lu, Marica Manisera, Xiaojun Mao, Yasumasa Matsuda, Kaiji Motegi, Jouchi Nakajima, Long Nguyen, Yoichi Nishiyama, David Nott, Teppei Ogihara, Hernando Ombao, Yasuhiro Omori, Alexander Petersen, Yasutaka Shimizu, Seung Jun Shin, Russell Shinohara, Zhihua Su, Yanlin Tang, Tiejun Tong, GuanNan Wang, Wendun Wang, Toshiaki Watanabe, Kenneth D West, Gongjun Xu, Jason Xu, Yuhang Xu, Weixin Yao, Danna Zhang, Jiwei Zhao, Yichuan Zhao, Yunpeng Zhao and Weixuan Zhu.

Local Organizing Committee:

Waseda University, EcoSta, CMStatistics and CFEnetwork.

Dear Colleagues,

It is a great pleasure to welcome you to the 6th International Conference on Econometrics and Statistics (EcoSta 2023). In light of the recent pandemic, we have introduced a hybrid format that accommodates both in-person and virtual attendance, ensuring flexibility for participants based on their circumstances and local restrictions. The conference program has been thoughtfully tailored to facilitate the optimal presentation of research findings and networking opportunities.

We are thrilled to host the largest gathering in the conference series, with approximately 1150 attendees and an impressive number of presentations. EcoSta 2023 comprises over 270 sessions, three keynote talks, four invited sessions, and around 1050 presentations. These figures serve as a testament to the support of our research communities, highlighting the significance of this initiative. We trust that the EcoSta conference will continue to serve as an excellent platform for disseminating high-quality research in Econometrics and Statistics while fostering valuable networking opportunities.

The conference is jointly organized by the Computational and Methodological Statistics Working Group (CMStatistics), the Computational and Financial Econometrics Network (CFEnetwork), the journal Econometrics and Statistics (EcoSta), and Waseda University. Building upon the achievements of previous editions, our aim is for this conference to become a leading event in the field of econometrics, statistics, and their applications.

The co-chairs express their gratitude to the scientific program committee, session organizers, and local organizing committee for their collective efforts in delivering a comprehensive program that covers various areas of econometrics and statistics. The Department of Applied Mathematics, Institute for Mathematical Science, Center for Data Science, Faculty of Commerce, Faculty of International Research and Education, and Faculty of Science and Engineering at Waseda University, as local hosts, along with the dedicated assistants, has played an instrumental role in ensuring the smooth organization of the conference. We extend our heartfelt thanks to all of them for their invaluable support.

The Elsevier journals of Econometrics and Statistics (EcoSta) and Computational Statistics & Data Analysis (CSDA) are associated with CFEnetwork, CMStatistics, and the EcoSta 2023 conference. The participants are encouraged to join the networks and submit their papers to special or regular peer-reviewed EcoSta and the CSDA Annals of Statistical Data Science (SDS) issues.

The inaugural **impact factor** for the journal Econometrics and Statistics (EcoSta) in 2022, announced in June 2023, stands at 1.9. Meanwhile, Computational Statistics & Data Analysis (CSDA) continues to maintain its commendable and consistent performance, with an impact factor of 1.8 for 2022.

Finally, we are happy to announce that the 7th International Conference on Econometrics and Statistics (EcoSta 2024) will take place at the Beijing Normal University, Beijing, China, 17-19 July 2024.

Ana Colubi, Erricos J. Kontoghiorghes and Yan Liu
on behalf of the Co-Chairs and EcoSta Editors

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 about 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 by 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 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. Now, 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 (Japan time, GMT+9)

2023-08-01

Opening, 08:35 - 08:50

A
EcoSta2023
08:50 - 09:40

Coffee break
09:40 - 10:10

B
EcoSta2023
10:10 - 11:50

Lunch break
11:50 - 13:20

C
EcoSta2023
13:20 - 15:00

D - Keynote
EcoSta2023
15:10 - 16:00

Coffee break
16:00 - 16:30

E
EcoSta2023
16:30 - 18:35

Welcome reception
18:35 - 19:30

2023-08-02

F
EcoSta2023
08:15 - 10:20

Coffee break
10:20 - 10:50

G
EcoSta2023
10:50 - 12:30

Lunch break
12:30 - 14:00

H
EcoSta2023
14:00 - 15:40

Coffee break
15:40 - 16:10

I
EcoSta2023
16:10 - 17:50

J
EcoSta2023
18:00 - 19:15

Conference dinner
19:30 - 21:30

2023-08-03

K
EcoSta2023
07:30 - 09:10

L
EcoSta2023
09:20 - 11:00

Coffee break
11:00 - 11:30

M - Keynote
EcoSta2023
11:30 - 12:20

Lunch break
12:20 - 13:40

N
EcoSta2023
13:40 - 14:55

Coffee break
14:55 - 15:25

O
EcoSta2023
15:25 - 17:05

P - Keynote
EcoSta2023
17:15 - 18:05

Closing, 18:05 - 18:20

TUTORIAL, MEETINGS, SOCIAL EVENTS AND ACCESS TO THE CONFERENCE

TUTORIAL

The tutorial “Confidence distribution: A new statistical inference approach and its applications in meta-analysis and fusion learning” will take place on Monday, 31st of July 2023, 15:00-19:30 (GMT+9) in the Room 201 (floor 2), Building 57, Nishi-Waseda Campus (see map on page VII), 3-chome-4 Okubo, Shinjuku City, Tokyo 169-0072, Japan. It will be delivered by Prof. Regina Liu, Rutgers, USA. Only participants who had subscribed for the tutorial can attend. Registered participants will be able to access the tutorial either in person or virtually.

SPECIAL MEETINGS by invitation to group members

The EcoSta (Econometrics and Statistics) and CSDA (Computational Statistics and Data Analysis) Editorial Board meetings will take place on Monday the 31st of July 2023, 17:30-18:15 (GMT+9). Indications to attend the Editorial Board meetings will be sent to the associate editors participating in the conference in due course.

ACCESS TO THE CONFERENCE

- Participants can attend virtually or in person accoring to what they selected while registering.
- The in-person venue is the Waseda University, 1 Chome-6 Nishiwaseda, Shinjuku City, Tokyo, Japan, Buildings 11 and 15 (see maps on page VIII). Please note that the university opens at 8:00, so you will not be able to enter before that time.
- The **registration**, **coffee breaks and lunches** will be located in the entrance (pilotis) of the ground floor of Building 11 (see maps on page VIII and floor maps on page IX).
- For environmental sustainability reasons, the conference endeavors to minimize paper usage and overall consumption. While there will be a limited number of printed Book of Abstracts, bags, pens, and pads available for those who require them, we strongly encourage all participants to opt for digital materials by downloading them onto their personal devices. For those who do utilize printed materials, we kindly request that they be returned after use to be reused or recycled. QR codes will be displayed in the registration area. These codes will enable participants to quickly access essential information, further reducing the need for printed materials and promoting a paperless conference experience.
- The conference is live streaming, and it will not be recorded. The oral presentations will take place through Zoom, while the virtual social events and poster presentations will run in Gather Town. The conference programme time is set at GMT+9.
- In order to access the virtual conference, you must first log in to the registration tool, get the daily password there and leave the session open. Then you should open another tab and go to the interactive programme (schedule). Click on the slot you wish to attend and then on the session. If it is hybrid, click on the blinking room of the floor map. You will be redirected to Zoom, where you will need to use the daily password.
- Please note that for security reasons, the Zoom links will not be sent to the speakers, and they can only be found on the online interactive programme (schedule).
- Detailed indications for virtual and in-person attendance, hybrid sessions, speakers, chairs, posters, networking, test sessions, as well as FAQ, can be found on the webpage.

Presentation instructions

The paper presentations must be shared through Zoom. The in-person rooms will be visible in Zoom as the corresponding hybrid session. Virtual speakers should install the application, have a stable internet connection, and make sure their video and audio work. They will share their slides when the chair requires it, present their talk, and be ready to answer the question after the presentation. Detailed indications for speakers can be found on the website. Each speaker has 20 minutes for the talk and 3-4 mins for discussion as a general rule. Strict timing must be observed.

Posters

The poster sessions will take place through Gather Town. The posters should be sent in **png format** to info@CMStatistics.org by the 27th July 2023. Landscape orientation is advisable. Detailed indications for the poster presentations can be found on the website.

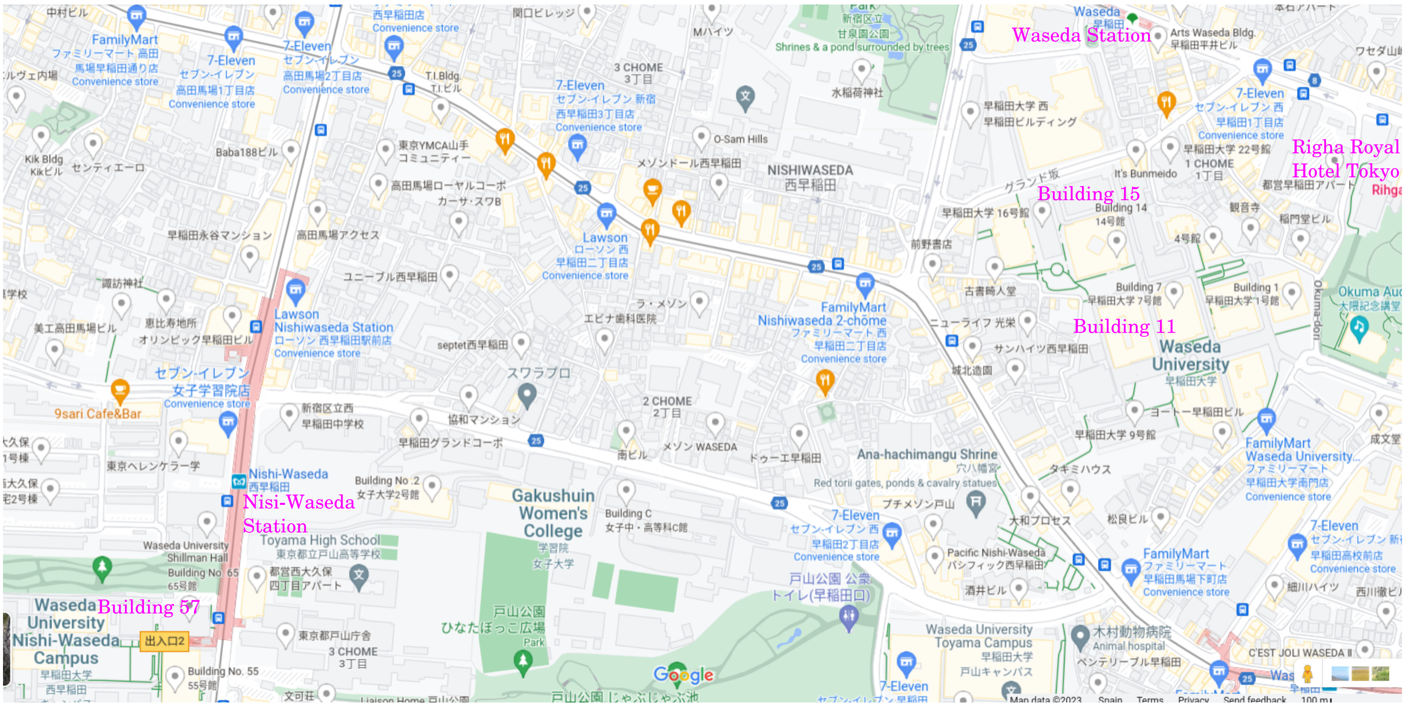
Session chairs

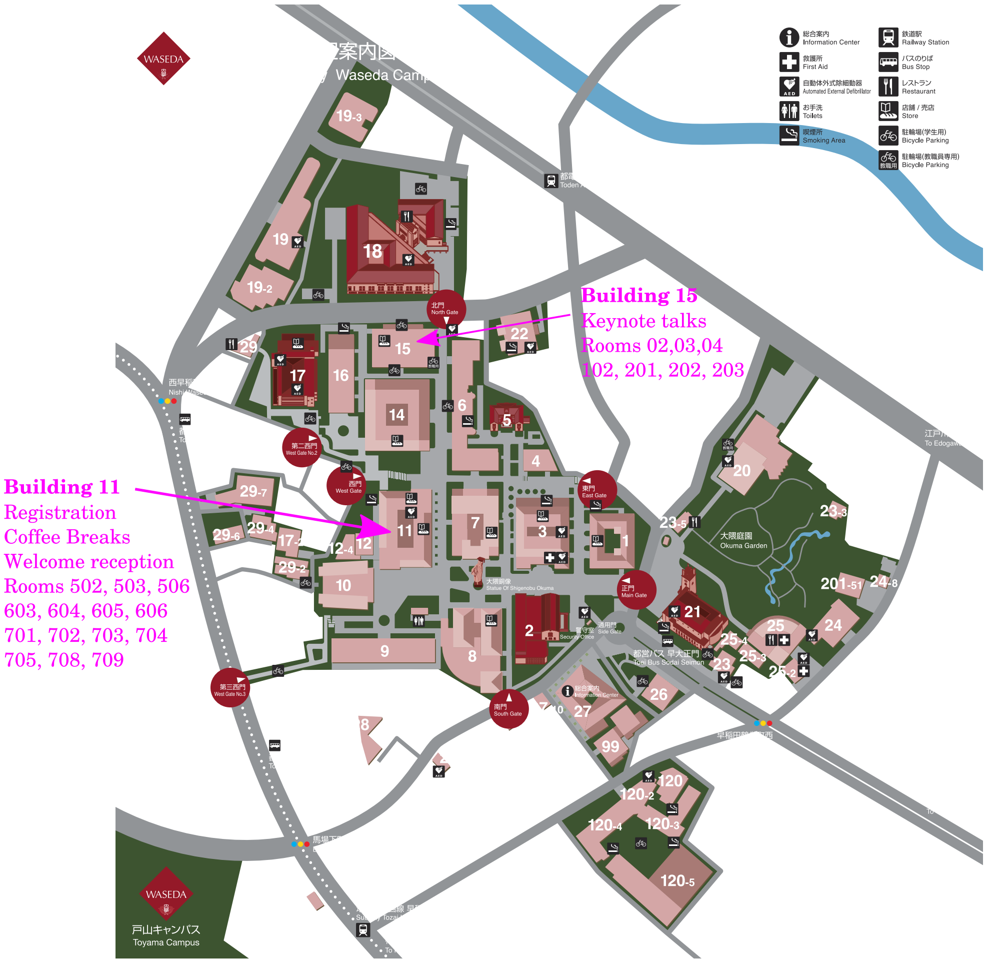
The session chairs will be responsible for introducing the session, the speakers and coordinating the discussion time. A member of the conference staff, identified by the name *Angel* followed by the room number, will assist in giving the rights to participate as the chair requests it. If any speaker is missing or has a technical problem, the chair can pass to the next speaker and come back later to resume if possible. Detailed indications for the (virtual or in-person) session chairs can be found on the website.

Test session

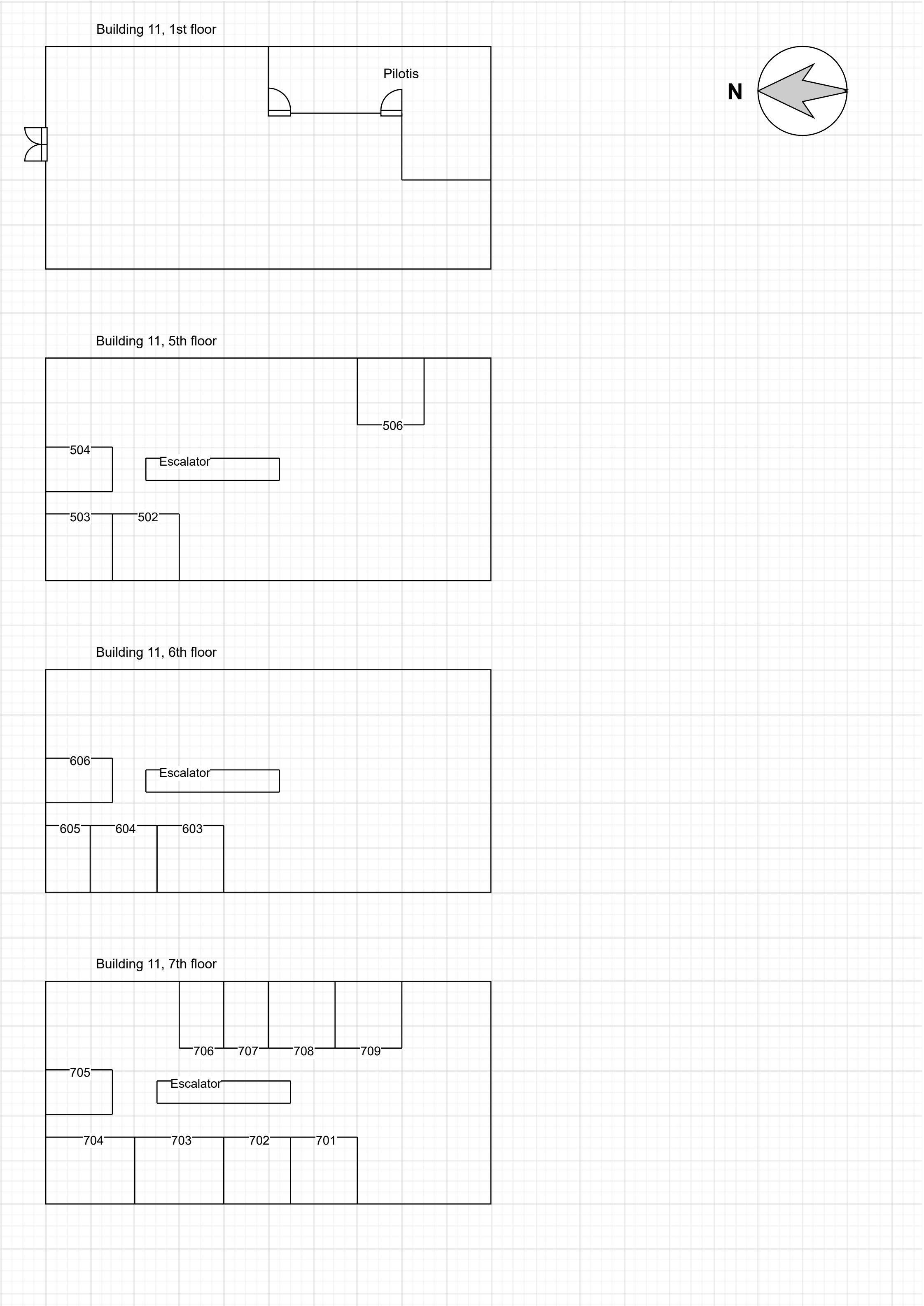
A test session will be set up for Saturday the 29th of July 2023, from 15:00 to 15:30 GMT+9 (Japan time). The participants will be able to enter through the Virtual Room R01 in the programme to test their presentations, (e.g., through Parallel session A) to test their presentations, video, micro and audio. Detailed indications for the test session can be found on the website.

Map of the venue and nearby area

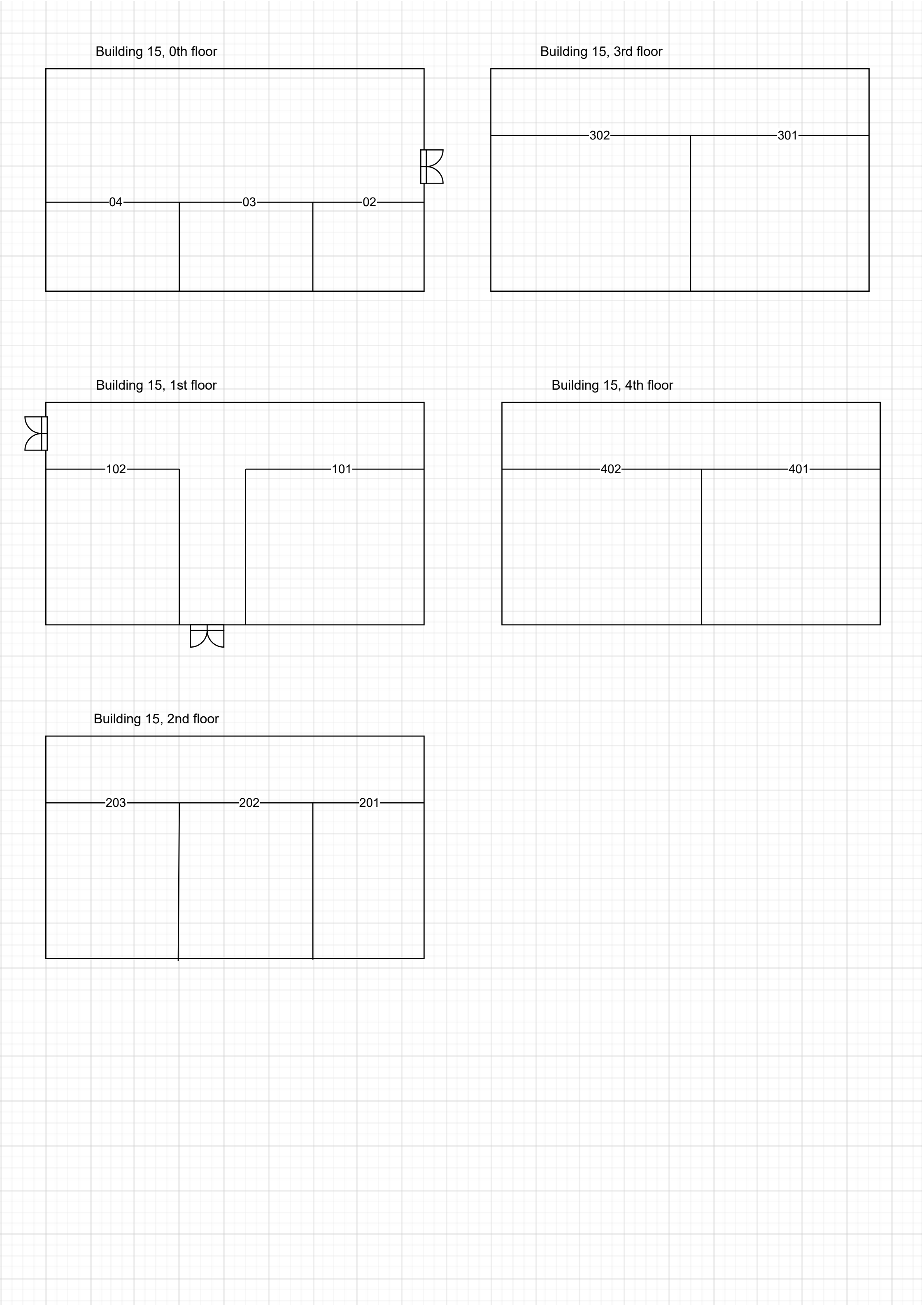




Floor maps



Building 11


Building 15, 0th floor
04
03
02
Building 15, 3rd floor
302
301
Building 15, 1st floor
102
101
Building 15, 4th floor
402
401
Building 15, 2nd floor
203
202
201

Building 15

PUBLICATION OUTLETS

The Elsevier journal Econometrics and Statistics (EcoSta) is the official journal of the conference. The CMStatistics network, co-organizer of the conference, also publishes the Annals of Statistical Data Science as a supplement to the journal Computational Statistics and Data Analysis (CSDA).

Econometrics and Statistics (EcoSta)

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

Econometrics and Statistics is the official journal of the networks Computational and Financial Econometrics and Computational and Methodological Statistics published by Elsevier (http://www.journals.elsevier.com/econometrics-and-statistics/). It publishes research papers in all aspects of econometrics and statistics and comprises of two sections:

- **Part A: Econometrics.** Emphasis will be given to methodological and theoretical papers containing substantial econometrics derivations or showing a potential of 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 to be considered when they involve an original methodology. Innovative papers in financial econometrics and its applications will be considered. The topics to be covered include portfolio allocation, option pricing, quantitative risk management, systemic risk and market microstructure. Interest will be 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 will 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 will not be of interest to the journal.

- **Part B: Statistics.** Papers providing important original contributions to methodological statistics inspired in applications will be considered for this section. Papers dealing, directly or indirectly, with computational and technical elements will be 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, review 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.

CSDA Annals of SDS

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

CMStatistics is inviting submissions for 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 theretical 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, 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 Editorial Manager system (choose Article Type: Research paper, and then Select “Section IV. Annals of Statistical Data Science”).

Contents

Tuesday 01.08.2023 15:10 - 16:00 Room: 102 Chair: Yan Liu Keynote talk 1

Fusion learning: Combining inferences from diverse data sources with heterogeneous data
Speaker: **Regina Liu, Rutgers University, United States** Dungang Liu, Min-ge Xie

Nowadays, advanced data collection technology makes inferences from diverse data sources easily accessible. Fusion learning combines inferences from multiple sources or studies to make more effective inferences than from any individual source or study alone. The tasks are focused on: 1) Whether/When to combine inferences? 2) How to combine inferences efficiently if you need to? A general framework for nonparametric and efficient fusion learning for inference on multi-parameters, which may be correlated, is present. The main tool underlying this framework is the new notion of depth confidence distribution (depth-CD), which is developed by combining data depth, bootstrap and confidence distributions. It is shown that depth-CD is an omnibus form of confidence region, whose contours of level sets shrink toward the true parameter value, and thus an all-encompassing inferential tool. The approach is shown to be efficient, general and robust. Specifically, it achieves high-order accuracy and Bahadur efficiency under suitably chosen combining elements. It readily applies to heterogeneous studies with a broad range of complex and irregular settings. This property also enables the approach to utilize indirect evidence from incomplete studies to gain efficiency for the overall inference.

Thursday 03.08.2023 11:30 - 12:20 Room: 102 Chair: Ana Colubi Keynote talk 2

Network approaches in healthcare, business and social sciences
Speaker: **Mike So, The Hong Kong University of Science and Technology, Hong Kong**

The use of network science to analyze complex relationships among variables or individuals has been recognized as an important approach in the literature. In statistical research, efforts have been devoted to developing network models, statistical inference procedures and prediction in time series of networks. Regarding applications, cases can be found in classical social network analysis, studies of interfirm relationships and systemic risk through the financial news, and the assessment of COVID-19 pandemic risk. Challenges in statistical modelling and inference are discussed. Recent examples of applying dynamic network ideas for tracking human behaviours and financial and pandemic risks are also reviewed. The examples also demonstrate the importance of interdisciplinary collaborations, including econometrics and statistical methods, in network science research for sustainable societal impacts.

Thursday 03.08.2023 17:15 - 18:05 Room: 102 Chair: Erricos Kontoghiorghes Keynote talk 3

Sufficient dimension reduction meets two-sample regression estimation
Speaker: **Masayuki Hirukawa, Ryukoku University, Japan**

When conducting regression analysis, econometricians often face situations where some regressors are unavailable in the dataset (e.g., an ability measure in wage regression). Suppose they can find an auxiliary dataset containing the missing regressors and several other variables common across two datasets. Previously, the problem of estimating regression parameters consistently by combining two datasets, proposing the matched-sample indirect inference (MSII) and plug-in least squares (PILS) estimators, respectively, was studied. However, these estimators can attain the parametric convergence rate only if the number of common variables is no greater than four. Then, under the assumption that the reduced form of each missing regressor can be expressed in a single-index form of the common variables, MSII and PILS are extended to overcome the curse of dimensionality. Restoring the parametric convergence rate for these estimators takes three steps, namely, (i) estimating index coefficients via some algorithms for sufficient dimension reduction; (ii) imputing proxies of the missing regressors; and (iii) estimating coefficients of the regression model. The convergence properties of these estimators are explored, and their finite-sample properties are examined via Monte Carlo simulations.

Tuesday 01.08.2023 08:50 - 09:40 Parallel Session A – EcoSta2023

EV316 Room Virtual R01 FUNCTIONAL DATA ANALYSIS Chair: Long Nguyen

E0263: **Regularized halfspace depth for functional data**
Presenter: **Hyemin Yeon**, Iowa 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 depths 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 a degeneracy issue. Here, 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 the centre, monotonicity relative to the deepest point, and upper semi-continuity, are established. Moreover, the proposed regularized halfspace depth can rank functional data with a varying emphasis in shape or magnitude, depending on the regularization. A new outlier detection approach is also proposed, which is capable of detecting both 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, our methods are shown to perform well in terms of detecting outliers of different types. Three real data examples showcase the proposed depth notion.

E0266: **A generalized functional linear model with spatial dependence**
Presenter: **Sooran Kim**, Iowa State University, United States
Co-authors: Xiongtao Dai, Mark Kaiser

A regression model is developed for spatially dependent binary response variables when the covariates form functional processes over time at each location for which the response is observed. The functional covariates are modelled in terms of a Fourier basis truncated to a finite number of terms. Responses are considered Markov random fields with conditional binary distributions and isotropic spatial dependence. Estimation is approached using a composite likelihood constructed from full conditional response distributions, sometimes also called Besags original pseudolikelihood in the spatial literature. Asymptotic properties are given for maximum composite likelihood estimators using a repeating lattice context, and the use of the model is illustrated with data relating new COVID vaccination rates in June for counties to the number of weekly infections reported over the previous several months in those same counties.

E0729: **Functional data driven financial risk management: An application to the NASDAQ Index**
Presenter: **Fatimah Alshahrani**, Princess Nourah bint Abdulrahman University, Saudi Arabia
Co-authors: Zoulikha Kaid, Zouaoui Chikr Elmezouar, Ali Laksaci, Raja M Almarzoqi

A new data-driven approach is introduced to manage financial risk in a real-time context. Inspired by the recent development of big data modelling, new dynamic models of financial risk analysis are developed using the statistical analysis of random curve data. More precisely, market risk is assessed through the Functional Expectile Regression (FER), the Functional Conditional Value at Risk (FCVR) and the Functional Conditional Expected Shortfall (FCES). These functional models are estimated using the nonparametric approach. Because the distribution of the volatility of the financial time series is usually not characterized by a specific distribution, the nonparametric approach is more adequate than the parametric for the financial time series. Thus, combining the functional smoothing of the financial data with the nonparametric adjustment of the risk models allows us to increase the accuracy of the existing risk models that are based on the parametric multivariate approach. This approach has been confirmed by an empirical analysis performed on NASDAQ composite index data. It covers 20 years of the daily return of the NASDAQ market index. A comparative study between the three functional models using different back-testing measures demonstrates that the FER model shows more variability, which allows better detection of the risk in crisis as well as in calm periods.

EI053 Room 604 ADVANCES IN ECONOMETRICS AND STATISTICS Chair: Donggyu Kim

E0163: **Fast inference for quantile regression with tens of millions of observations**
Presenter: **Myung Hwan Seo**, Seoul National University, Korea, South
Co-authors: Yuan Liao, Youngki Shin, Sokbae Lee

Big data analytics has opened new avenues in economic research, but the challenge of analyzing datasets with tens of millions of observations is substantial. Conventional econometric methods based on extreme estimators require large amounts of computing resources and memory, often not readily available. It is focused on linear quantile regression applied to ultra-large datasets, such as U.S. decennial censuses. A fast inference framework utilizing stochastic sub-gradient descent (S-subGD) updates is presented. The proposed test statistic is calculated fully online, and critical values are calculated without resampling. Extensive numerical studies are conducted to showcase the computational merits of the proposed inference. For inference problems as large as (n,d)(107,103), where n is the sample size and d is the number of regressors, the method generates new insights, surpassing current inference methods in computation. The method specifically reveals trends in the gender gap in the U.S. college wage premium using millions of observations while controlling over 103 covariates to mitigate confounding effects.

E0164: **Reinforcement learning via nonparametric smoothing in a continuous-time stochastic setting**
Presenter: **Shang Wu**, Fudan University, China
Co-authors: Yazhen Wang

Reinforcement learning is mainly developed for discrete-time Markov decision processes. A method is proposed to establish a novel learning approach based on temporal difference and non-parametric smoothing to solve reinforcement learning problems in a continuous-time stochastic setting. Continuous-time temporal-difference learning developed for deterministic models is unstable and diverges when applied to data generated from stochastic models. It is shown that the proposed learning approach has a robust performance for data generated from stochastic models governed by stochastic differential equations. The asymptotic theory is established for the proposed approach, and a numerical study is carried out to solve a pendulum reinforcement learning problem and check the finite sample performance of the proposed method.

E0890: **Identifying the effect of persuasion**
Presenter: **Sokbae Lee**, Columbia University, United States
Co-authors: Sung Jae Jun

A commonly used measure of persuasion is examined, whose precise interpretation has been obscure in the literature. Using the potential outcome framework, the causal persuasion rate is defined by a proper conditional probability of taking the action of interest with a persuasive message conditional on not taking action without the notification. Identification is then formally studied under empirically relevant data scenarios. The commonly adopted measure generally that does not estimate is shown, but the causal rate of persuasion is often overstated. Several new parameters of interest are discussed, and practical methods for causal inference are provided.

EO096 Room 02 PERSONALIZED DECISION-MAKING WITH LONGITUDINAL DATA Chair: Yifan Cui

E1275: **Post-episodic reinforcement learning inference**
Presenter: **Ruohan Zhan**, Hong Kong University of Science and Technology, Hong Kong
Co-authors: Vasilis Syrgkanis

Estimation and inference with data collected from episodic reinforcement learning (RL) algorithms are considered, i.e. adaptive experimentation algorithms that at each period (aka episode) sequentially interact multiple times with a single treated unit. The goal is to evaluate counterfactual adaptive policies after data collection and estimate structural parameters such as dynamic treatment effects, which can be used for credit assignment (e.g., what was the effect of the first-period action on the final outcome). Such parameters of interest can be framed as solutions to moment equations but not minimizers of a population loss function, leading to Z-estimation approaches in the case of static data. However, such estimators fail to be asymptotically normal in the case of adaptive data collection. A re-weighted Z-estimation approach is proposed with carefully designed adaptive weights to stabilize the episode-varying estimation variance resulting from the nonstationary policy typical episodic RL algorithms invoke. Proper weighting schemes are identified to restore the consistency and asymptotic normality of the re-weighted Z-estimators for target parameters, which allows for hypothesis testing and constructing reliable confidence regions for target parameters of interest. Primary applications include dynamic treatment effect estimation and dynamic off-policy evaluation.

E1291: **Dynamic assortment selection with position effects**
Presenter: **Yiyun Luo**, School of Statistics and Management, Shanghai University of Finance and Economics, China

In online retailing and advertising, the seller aims to offer the customers an assortment of items that incur maximal expected revenue. A new online decision-making problem, Dynamic Assortment Selection with Positioning (DAP)d, is proposed. There are two key characteristics of the DAP problem. Firstly, customers' preferences for the items are unknown at the beginning and thus need to be learned from interactions with customers. Secondly, the selected assortment and the positioning of items in the assortment would influence the customers' purchasing behaviours and, thus, the revenues. The goal is maximising overall revenue in a finite horizon by making sequential assortment and positioning decisions. Specifically, for the DAP problem, a UCB-based policy with sublinear regrets is developed. The proposed policy delivers superior performances in various simulation settings by handling the position effects well.

E1301: **Towards trustworthy explanation: On causal rationalization**
Presenter: **Hengrui Cai**, University of California Irvine, United States

With recent advances in natural language processing, rationalization becomes an essential self-explaining diagram to disentangle the black box by selecting a subset of input texts to account for the major variation in prediction. Yet, existing association-based approaches on rationalization cannot identify true rationales when two or more snippets are highly inter-correlated and thus provide a similar contribution to prediction accuracy, so-called spuriousness. To address this limitation, two causal desiderata, non-spuriousness and efficiency, are novelly leveraged into rationalization from the causal inference perspective. A series of probabilities of causation is formally defined based on a newly proposed structural causal model of rationalization, with its theoretical identification established as the main component of learning necessary and sufficient rationales. The superior performance of the proposed causal rationalization is demonstrated on real-world review and medical datasets with extensive experiments compared to state-of-the-art methods.

EO011 Room 03 MODERN ANALYTICAL METHODS FOR BIOMEDICAL RESEARCH Chair: Jiwei Zhao

E0944: **Learning individualized minimal clinically important difference (iMCID) from high-dimensional data**
Presenter: **Jiwei Zhao**, University of Wisconsin-Madison, United States

Statistical significance has been widely used to infer the treatment effect in assessing the efficacy of a treatment or intervention; however, there has been a growing recognition that statistical significance has limitations. On the contrary, clinical significance is usually desirable in practice as it provides a better assessment of clinically meaningful improvement. A critical concept in evaluating clinical significance is a minimal clinically important difference (MCID), the smallest change in the outcome that an individual patient would identify as important. A statistical learning framework for estimating the individualized MCID (iMCID) from high-dimensional data will be presented. In particular, a path-following iterative algorithm and some novel nonregular theoretical results will be presented. Additionally, simulation studies that reinforce the theoretical findings and an application to the study of chondral lesions in knee surgery to demonstrate the usefulness of the proposed approach will also be discussed.

E1295: **Imaging mediation analysis for longitudinal outcomes**
Presenter: **Cai Li**, St. Jude Children's Research Hospital, United States

The focus is improving cognitive outcomes for pediatric cancer survivors undergoing aggressive cancer treatments that may affect the central nervous system. Specifically, a new mediation model is proposed for longitudinal outcomes to a clinical trial for medulloblastoma, using high-dimensional imaging mediators to identify causal pathways and corresponding white matter microstructures. The proposed approach accounts for spatial dependency and smoothness of the mediators, improving the detection power of informative voxels. The results provide guidance on improving long-term neurodevelopment and sparing brain regions impacted by treatment. The validity of the method is confirmed through simulation studies.

E1067: **Regional variation of length of stay in inpatient rehabilitation facilities and the market-level influencing factors**
Presenter: **Jessica Cao**, University of Wisconsin-Madison, United States

The aim is to 1) measure the regional variation of length of stay (LOS) for post-acute care delivered in inpatient rehabilitation facilities (IRF); and 2) identify market-level factors and quantify their effects on regional variation. A multi-year cross-sectional design is adopted, and data from the Uniform Data System for Medical Rehabilitation (UDSMR) are used. The sample included IRF admissions for US Medicare beneficiaries 65 years or older with stroke as a primary diagnosis from January 2019 to December 2020. The primary outcomes are regional variations by Medicare coverage type (TM vs MA) and pandemic timing (before vs during). Cross-region variation is calculated as the variance of average risk-adjusted LOS for 10 CMS regions. Within-region variation is calculated as the variance of risk-adjusted LOS for all admissions within a region. The main explanatory variables included regional-level Medicare Advantage penetration and IRF market consolidation measured by Herfindahl Index (HHI). Risk adjustment included sociodemographic variables, clinical conditions, and facility characteristics. Results suggested that administrative variations in MA plans contribute to regional variation in LOS for MA-covered admissions, especially in regions with high MA penetration or low IRF market consolidation. Uniform administration and market consolidation can help reduce regional variation.

EO204 Room 04 ADVANCED AND PRACTICAL BAYESIAN METHODS FOR BIOMEDICAL RESEARCH Chair: Menggang Yu

E0425: **A uniform shrinkage prior in spatiotemporal Poisson models for count data**
Presenter: **Yisheng Li**, The University of Texas MD Anderson Cancer Center, United States

Default Bayesian inference in a Poisson generalized linear mixed model is considered for spatiotemporal data. Normal random effects are used to model the within-area correlation over time, and spatial effects represented with a proper conditional autoregressive (CAR) model are used to model the between-area correlations. A uniform shrinkage prior (USP) for the variance components of the spatiotemporal random effects is developed. It is proven that the proposed USP is proper, and the resulting posterior is proper under the proposed USP, an independent flat prior for each fixed effect, and a uniform prior for a spatial parameter under suitable conditions. Posterior simulation is implemented, and inference is made using the OpenBUGS, R2OpenBUGS, and RStan software packages. We illustrate the proposed method by applying it to a leptospirosis count dataset with observations from 17 northern provinces of Thailand across four quarters in 2011 to construct the disease maps. According to the deviance information criterion, the proposed USP for the variance components of the spatiotemporal effects yields better performance than the conventional inverse gamma priors. A simulation study suggests that the estimated fixed-effect parameters are accurate based on a relative bias criterion.

E1164: **External control designs to incorporate real-world evidence with adaptive information borrowing**
Presenter: **Jun Yin**, Mayo Clinic, United States
Co-authors: Peter Noseworthy, Xiaoxi Yao

Randomized clinical trials (RCTs) are considered the standard approach for assessing drug efficacy; however, patient recruitment in RCTs can be challenging for rare diseases. Innovative approaches can enhance clinical trials' efficiency, such as information borrowing, which involves leveraging information from external data (i.e., previously completed trials or contemporaneous practice data). A novel Bayesian method of information borrowing is proposed from external data in clinical trials, motivated by a Mayo Clinic pragmatic trial with real-world controls. The proposed method allows borrowing when appropriate but also accommodates heterogeneous scenarios. Performance is evaluated using simulation studies.

E0748: **Statistical inference of selection coefficient from temporal allele frequencies**
Presenter: **Yuehao Xu**, Johannes Kepler University Linz, Austria

The focus is on estimating selection coefficients for a Wright-Fisher Model. Given the inherently intractable nature of the model's likelihood function, the research delves into various approximation techniques to improve the accuracy and efficiency of selection coefficient estimations. Adhering to the Bayesian paradigm, posterior distributions and evaluate uncertainty are constructed in the results. Furthermore, innovative methods are incorporated to summarize high-dimensional allele frequency data during the likelihood approximation process. The methods aim to provide valuable insights into evolutionary mechanisms and aid in developing more accurate models for estimating the parameters of evolutionary forces over time. Compared to existing methods, the novel approach exhibits promising performance on both simulated and real-world data sets.

EO051 Room Virtual R02 RECENT ADVANCES IN ANALYSIS OF POINT PROCESS DATA Chair: Kuang-Yao Lee

E0342: **Latent network structure learning from high-dimensional multivariate point processes**
Presenter: **Emma Jingfei Zhang**, Emory University, United States

Learning the latent network structure from large-scale multivariate point process data is important in various scientific and business applications. For instance, it might be wished to estimate the neuronal functional connectivity network based on spiking times recorded from a collection of neurons. To characterize the complex processes underlying the observed data, a new and flexible class of nonstationary Hawkes processes is proposed that allow both excitatory and inhibitory effects. An efficient sparse least squares estimation approach estimates the latent network structure. Using a thinning representation, concentration inequalities are established for the first and second-order statistics of the proposed Hawkes process. Such theoretical results enable us to establish the non-asymptotic error bound and the selection consistency of the estimated parameters. Furthermore, a least squares loss-based statistic is described for testing if the background intensity is constant in time. The efficacy of the proposed method through simulation studies and an application is demonstrated to a neuron spike train dataset.

E0381: **Learning human activity patterns using clustered point processes with active and inactive states**
Presenter: **Biao Cai**, Yale University, United States

Modelling event patterns is a central task in a wide range of disciplines. In applications such as studying human activity patterns, events often arrive clustered with sporadic and long periods of inactivity. Such heterogeneity in event patterns poses challenges for existing point process models. A new class of clustered point processes is proposed that alternate between active and inactive states. The proposed model is flexible, highly interpretable, and can provide useful insights into event patterns. A composite likelihood approach and a composite EM estimation procedure are developed for efficient and numerically stable parameter estimation. Both the computational and statistical properties of the estimator, including convergence, consistency, and asymptotic normality, are studied. The proposed method is applied to Donald Trumps Twitter data to investigate if and how his behaviours evolved before, during, and after the presidential campaign. Additionally, large-scale social media data is analyzed from Sina Weibo and identify interesting groups of users with distinct behaviours.

E0423: **Generalized functional linear model with a point process predictor**
Presenter: **Jiehuan Sun**, University of Illinois at Chicago, United States
Co-authors: Kuang-Yao Lee

Point process data have become increasingly popular these days. For example, many of the data captured in electronic health records (EHR) are in the format of point-process data. It is of great interest to study the association between a point process predictor and a scalar response using generalized functional linear regression models. Various generalized functional linear regression models have been developed under different settings in the past decades. However, existing methods can only deal with functional or longitudinal predictors, not point-process predictors. A novel generalized functional linear regression model is proposed for a point process predictor. The proposed model is based on the joint modelling framework, where a log-Gaussian Cox process model is adopted for the point process predictor and a generalized linear regression model for the outcome. A new algorithm is also developed for fast model estimation based on the Gaussian variational approximation method. Extensive simulation studies are conducted to evaluate the performance of the proposed method and it is compared to competing methods. The performance of the proposed method is further demonstrated on an EHR dataset of patients admitted into the intensive care units of the Beth Israel Deaconess Medical Center between 2001 and 2008.

EO073 Room 102 SPATIAL EPIDEMIOLOGY Chair: Pei-Sheng Lin

E0331: **Regularized spatial and spatio-temporal cluster detection: Applications to breast cancer**
Presenter: **Maria Kamenetsky**, University of Wisconsin-Madison, United States
Co-authors: Jun Zhu, Ronald Gangnon, Junho Lee

There are patterns in how people and disease group across space and time. These patterns are important to epidemiologists and health professionals because they may indicate elevated disease risk. In some cases, this high risk may be driven by external factors such as environmental exposures, infectious diseases, changes in lifestyle factors or other factors where a timely public health intervention may save lives. The detection of disease clusters has typically been approached as a large-scale multiple testing problem using a spatial and spatio-temporal scan statistic. Instead, spatial cluster detection has been re-examined as a high-dimensional variable selection problem using (quasi-)Poisson regression penalized by the least absolute shrinkage and selection operator (LASSO). Using sparse matrices, fast and efficient computation is made possible by exploiting the effects of potential clusters. Final models are selected based on (quasi-)information criteria, which allows us to smooth over the background rate and identify the selected breast cancer clusters. Data-driven simulation results demonstrate our approach for detecting single and multiple spatio-temporal clusters. Practical applications of the methods are illustrated using data on breast cancer incidence in Japan.

E0616: **Understanding the spread of infectious diseases in edge areas of hotspots**
Presenter: **Tzai-Hung Wen**, National Taiwan University, Taiwan

Hotspots have been shown to have a higher potential for transmission risk, making them a priority for controlling epidemics. However, the role of edge areas of hotspots in disease transmission remains unclear. The aim is to examine whether disease incidence rate growth is higher on the edges of disease hotspots during outbreaks. Our data is based on Taiwan's three most severe dengue epidemic years from 1998 to 2020. Conditional autoregressive models and Bayesian areal Wombling methods are employed to identify significant edge areas of hotspots based on the extent of risk difference between adjacent areas. The difference-in-difference estimator in CAR models measures the growth rate of risk by comparing the incidence rate between two groups (hotspots and edge areas) over two time periods. The results show that the edge areas of hotspots have a more significant increase in disease risk than hotspots, leading to a higher risk of disease transmission and potential disease foci. This finding explains the diffusion mechanism of epidemics, a pattern mixed with expansion and relocation, indicating that the edge areas play an important role. The study highlights the importance of considering edge areas of hotspots in disease transmission.

E0918: **Minibus connectivity to brothels in Lilongwe, Malawi, and implications for STI clinic attendance**
Presenter: **Feng-Chang Lin**, University of North Carolina - Chapel Hill, United States
Co-authors: Skyler Noble, Griffin Bell

Social and geospatial connectivity often shapes the sexual networks through which sexually transmitted infections (STIs) propagate. Public transportation can play an important role in STI transmission. It bridges communities and facilitates shopping, medical visits, social outings, and even meeting potential sexual partners. Minibus transportation is common in Lilongwe, Malawi. However, its role in connecting residences to venues where sex is purchased and sold has not been explored. Minibus routes and brothel locations in Lilongwe are collected, and the connectivity is determined to these brothels via minibus routes. A generalized additive model (GAM) with a nonparametric spatial dependence was fit to explore the relative risk of STI clinic visits and the travelling time via the minibus. The rapid decay of the relative risks in travelling time shows the strong association between the connectivity of the transition network and the population's sexual health outcome.

EO077 Room 201 NONPARAMETRIC CAUSAL INFERENCE Chair: Nilanjana Laha

E0231: **Efficient estimation of modified treatment policy effects based on the generalized propensity score**
Presenter: **Nima Hejazi**, Harvard T.H. Chan School of Public Health, United States
Co-authors: Ivan Diaz, David Benkeser, Mark van der Laan

Continuous treatments have posed a significant challenge for causal inference, both in formulating and identifying scientifically meaningful effects and in their robust estimation. Traditionally, the focus has been placed on techniques applicable to binary or categorical treatments with few levels, allowing for applying propensity score-based methodology with relative ease. Efforts to accommodate continuous treatments introduced the generalized propensity score. Yet, estimators of this nuisance parameter commonly rely upon parametric regression strategies that sharply limit the robustness and efficiency of inverse probability weighted (IPW) estimators of causal effect parameters. A flexible generalized propensity score estimator with rate-convergence properties desirable in semiparametric theory is formulated. With this estimator, nonparametric IPW estimators of a class of causal effect estimands tailored to continuous treatments are constructed. To obtain asymptotic efficiency for the proposed estimators, several non-restrictive selection procedures are outlined for applying a sieve estimation framework to under smooth generalized propensity score estimators. In numerical experiments, these novel, nonparametric IPW estimators are demonstrated capable of achieving the nonparametric efficiency bound (comparable to so-called double robust estimators) in a setting with continuous treatments, and their higher-order efficiency properties are investigated.

E0867: **New root-n consistent, numerically stable higher-order influence function estimators**
Presenter: **Lin Liu**, Shanghai Jiao Tong University, China

Higher-Order Influence Functions (HOIFs) provide a unified theory for constructing rate-optimal estimators for a large class of low-dimensional (smooth) statistical functionals/parameters (and sometimes even infinite-dimensional functions) that arise in substantive fields, including epidemiology, economics, and the social sciences. Since introducing HOIFs, they have been viewed mainly as a theoretical benchmark rather than a valuable tool for statistical practice. Works aimed to flip the script are scant, but a few recent papers make some partial progress. A fresh attempt at achieving this goal by constructing new, numerically stable HOIF estimators (or sHOIF estimators for short, with s standing for stable) is taken with provable statistical, numerical, and computational guarantees. This new class of sHOIF estimators (up to the 2nd order) was foreshadowed in synthetic experiments conducted by other researchers.

E1266: **On statistical inference with high dimensional sparse CCA**
Presenter: **Nilanjana Laha**, Texas A&M University, United States

The focus is on asymptotically exact inference on the leading canonical correlation directions and strengths between two high-dimensional vectors under sparsity restrictions. The main contribution is developing a novel representation of the CCA problem, based on which one can operationalize a one-step bias correction on reasonable initial estimators. The analytic results are adaptive over suitable structural restrictions of the high dimensional nuisance parameters, which, in this setup, correspond to the covariance matrices of the variables of interest. The theoretical guarantees behind the procedures with extensive numerical studies are further supplemented.

EO165 **Room 203** **RECENT STATISTICAL ADVANCES IN BIOMEDICAL SCIENCES** **Chair: Binyan Jiang**

E0635: **Zero-inflated smoothing spline for single-cell data**
Presenter: **Xiaoxiao Sun**, University of Arizona, United States

Trajectory inference methods aim to order single cells along a trajectory based on their gene expression pattern in single-cell data. Such analyses provide new opportunities to investigate cellular dynamic processes such as cell cycle and cell differentiation, which are critical to study disease progression. However, due to the low capturing and sequencing efficiency of single-cell sequencing techniques, there is a common dropout issue, referring to the presence of excessive zero counts in the data. This dropout issue poses a challenge for traditional smoothing spline techniques to analyze single-cell trajectory data. To address this, a zero-inflated smoothing spline method specifically has been developed for single-cell trajectory data.

E0643: **Robust inference for federated meta-learning**
Presenter: **Xiudi Li**, Harvard University, United States
Co-authors: Zijian Guo

Synthesizing information from multiple data sources is critical to ensure knowledge gen- eralizability. Integrative analysis of multi-source data is challenging due to the heterogeneity across sources and data-sharing constraints due to privacy concerns. A general robust inference framework is considered for federated meta-learning of data from multiple sites, enabling statistical inference for the prevailing model, defined as the one matching the majority of the sites. Statistical inference for the prevailing model is challenging since it requires a data-adaptive mechanism to select eligible sites and subsequently accounts for the selection uncertainty. A novel sampling method is proposed to address the additional variation arising from the selection. The devised CI construction does not require sites to share individual-level data and is shown to be valid without requiring the selection of eligible sites to be error-free. The proposed robust inference for federated meta-learning (RIFL) methodology is broadly applicable and illustrated with three inference problems: aggregation of parametric models, high-dimensional prediction models, and inference for average treatment effects. RIFL is used to perform federated learning of mortality risk for patients hospitalized with COVID-19 using real-world EHR data from 16 healthcare centres representing 275 hospitals across four countries.

E0669: **Design of network-based studies to estimate individual and spillover effects and identify key influencer**
Presenter: **Zhibing He**, Yale University, United States

Many interventions in public health act in settings where individuals are connected to one another, and the intervention assigned to randomly selected individuals may spill over to their network members. The effects of such interventions can be quantified in several ways. The overall effect measures the average intervention effect across the study population over those directly treated, along with those to whom the intervention spills over but who are not directly treated. The individual effect measures the intervention effect among those directly treated, while the spillover effect measures the effect among those in the network of those directly treated. Here, methods for study design with the aim of estimating individual, spillover and overall effects are developed. In addition, new study designs to find the most influential participants are developed by identifying the heterogeneity of the spillover effect. In particular, an ego network-based randomized design is developed in which a set of index participants is sampled from the population and randomly assigned to treatment while data are also collected for their untreated network members.

EO018 **Room 503** **ADVANCES IN NONPARAMETRIC INFERENCE, FAIRNESS, AND IV REGRESSION** **Chair: Daoji Li**

E0998: **Estimating and implementing conventional fairness metrics with probabilistic protected features**
Presenter: **Patrick Vossler**, Stanford University, United States
Co-authors: Hadi Elzayn, Emily Black, Nathan Jo

Techniques from algorithmic fairness are increasingly used to train models that satisfy various quantitative notions of fairness, hoping to avoid potential bad outcomes observed in various machine learning-based settings. Yet the vast majority of such techniques require access to the protected attribute, either at train time or in production, and this protected attribute is often unavailable. It is shown how to leverage probabilistic race imputation, most notably via Bayesian Improved Surname Geocoding (BISG), along with special estimators, to estimate quantitative fairness measurements without access to the protected feature. Under certain conditions described, these estimators will be unbiased; otherwise, they can be used as bounds. This technique can be used both simply to document or rule out unfairness (according to the particular definition) when given a model. It can be incorporated into many techniques designed to produce fair models. Quantitative guarantees are proved for the proposed methods, and two empirical illustrations are provided based on tax data from the IRS and medical data from AFC.

E1148: **Optimal invariant tests in an instrumental variables regression with heteroskedastic and autocorrelated**
Presenter: **Mahrad Sharifvaghefi**, University of Pittsburgh, United States

Model symmetries in the instrumental variable (IV) regression are used to derive an invariant test for the causal structural parameter. Contrary to popular belief, it is shown that there exist model symmetries when equation errors are heteroskedastic and autocorrelated (HAC). The theory is consistent with existing results for the homoskedastic model. These symmetries are used to propose the conditional integrated likelihood (CIL) test for the causality parameter in the over-identified model. Theoretical and numerical findings show that the CIL test performs well compared to other tests in terms of power and implementation. Practitioners use the Anderson-Rubin (AR) test in the just-identified model, and the CIL test in the over-identified model is recommended.

E0953: **Improvement of gradient boosting using regularization and optimization algorithms**
Presenter: **Hideo Suzuki**, Keio University, Japan
Co-authors: Nagomu Iwasa

Several methods are proposed for gradient boosting by introducing regularization and optimization algorithms, such as Momentum SGD, Adadelta, and Adam. Regularization has the effect of suppressing overfitting by constraining the degrees of freedom of the constructed model. The regularization term is calculated numerically and added to the loss function obtained from the residual between the predicted and measured values. The L1 and L2 regularization terms in the scores of all decision tree leaves are used. In the conventional SGD, the training data is shuffled, one is randomly extracted from it, the error is calculated, and the parameters are updated to reduce the loss function using the gradient method. The optimization algorithms, which are improved versions of the conventional SGD, suppress vibration by using the gradient information from the previous period, which enables us to alleviate the problems of the conventional SGD. To verify the effect of regularization and improved algorithms on gradient boosting, the predictive accuracy and calculation efficiency indicators for several datasets of the UCI Machine Learning Repository are measured, and those of the conventional SGD, SGD (regularization), SGD (improved algorithms) and SGD (regularization +improved algorithms) are compared. The result shows that SGD (regularization+Adam) is generally good regarding prediction accuracy and calculation efficiency.

E1334: **Optimal nonparametric inference with two-scale distributional nearest neighbors**
Presenter: **Emre Demirkaya**, University of Tennesse, Knoxville, United States
Co-authors: Lan Gao, Jinchi Lv, Yingying Fan, Jingbo Wang, Patrick Vossler

The weighted nearest neighbours (WNN) estimator has been popularly used as a flexible and easy-to-implement nonparametric tool for mean regression estimation. The bagging technique is an elegant way to form WNN estimators with weights automatically generated to the nearest neighbours; the resulting estimator as the distributional nearest neighbours (DNN) for easy reference is named. Yet, there is a lack of distributional results for such an estimator, limiting its application to statistical inference. Moreover, when the mean regression function has higher-order smoothness, DNN does not achieve the optimal nonparametric convergence rate, mainly because of the bias issue. An in-depth technical analysis of the DNN is provided, based on which a bias reduction approach for the DNN estimator is suggested by linearly combining two DNN estimators with different subsampling scales, resulting in the novel two-scale DNN (TDNN) estimator. The two-scale DNN estimator is proven to enjoy the optimal nonparametric convergence rate in estimating the regression function under the fourth-order smoothness condition. Further beyond estimation, it was established that the DNN and two-scale DNN are asymptotically normal as the subsampling scales and sample size diverge to infinity. The theoretical results and appealing finite-sample performance of the suggested two-scale DNN method are illustrated with simulation examples and a real data application.

EO032 Room 506 ADVANCED STATISTICAL LEARNING APPROACHES IN ANALYZING COMPLEX MODERN DATA Chair: Xiwei Tang

E0474: **Innovative precision medicine methods in subgroup identification for Alzheimers disease**
Presenter: **Lei Liu**, Washington University in St. Louis, United States

Alzheimer's disease (AD) is a progressive, degenerative disorder of the brain and is the most common form of dementia of ageing. Developing new and more effective medications to treat Alzheimer's disease remains a high priority. However, considerable heterogeneity exists among people with Alzheimer's disease, which might affect their reactions to medications differently. To improve the safety, efficacy, and efficiency of Alzheimer's treatment medications, it is important to identify and treat patients who are likely to respond best to a particular medication. Precision medicine, which uses individual features to diagnose and treat disease, is of growing interest in Alzheimer's treatment. The aim is to develop new machine-learning methods to select predictive biomarkers and identify subgroups of patients who showed an enhanced treatment effect in a recently completed AD trial. The predictive biomarkers can identify patients more likely to benefit from treatment. Subgroups will be found by defining a combination of multiple predictive biomarkers through their interactions with treatment.

E0632: **Recent advances in fast estimation of high-dimensional hidden Markov models**
Presenter: **Jordan Rodu**, University of Virginia, United States
Co-authors: Jordan Rodu, Xiaoyuan Ma

Hidden Markov models (HMMs) are powerful models commonly used for modelling time series. HMMs are typically estimated using a special case of the E-M algorithm called the Baum-Welch algorithm. However, the Baum-Welch algorithm can be slow, is prone to finding local optima, and can struggle (or even outright fail) with high-dimensional data. An alternative approach to estimating HMMs, called spectral estimation of HMMs (sHMM), which overcomes these challenges, is discussed. While sHMM is ideal for real-time, high-dimensional scenarios where fast estimation and adaptation are crucial, it can suffer from instability in some cases. Recent developments that address this issue are highlighted, making sHMM a crucial tool in the time series toolbox. Examples from high-frequency trading are provided.

E0704: **A multi-use graph neural network framework for single-cell multi-omics data**
Presenter: **Peifeng Ruan**, UT Southwestern Medical Center, United States

The advances of single-cell multi-omics profiling technologies in biomedical research offer an unprecedented opportunity for understanding cell heterogeneity and subpopulations. There are many statistical and computational challenges in the integrative analyses of these rich data, including sequencing sparsity, complex differential patterns in gene expression, and different platforms and panels used to generate multiple single-cell multi-omics. A multi-use graph neural network framework is introduced that can effectively impute, and missing sequencing panels are predicted, multi-omics single-cell datasets are integrated, and cell-cell relationships with graph neural networks are formulated and aggregated. Comprehensive simulations and applications on multiple CITE-seq and single-cell RNA-sequencing datasets demonstrate that the proposed method is a powerful tool for general single-cell data multi-omics analyses that outperforms the existing methods for protein prediction, gene imputation and cell clustering.

EO112 Room 603 NONPARAMETRIC METHODS IN FINANCE AND ECONOMETRICS Chair: Huanjun Zhu

E1304: **Bayesian nonparametric portfolio selection with rolling maximum drawdown control**
Presenter: **Mei Xiaoling**, Xiamen University, China

The portfolio selection problem is considered for a multiperiod investor who seeks to maximize mean-variance utility facing multiple risky assets and various trading constraints in the presence of return predictability. With the presence of trading constraints, dynamic programming is impractical to carry out due to the curse of dimensionality. Model predictive control is implemented, which is computationally efficient to solve the problem, and it is proposed to use a nonparametric Bayesian model, i.e., hierarchical Dirichlet process-based Hidden Markov Model (HDP-HMM) to predict the multiperiod mean and covariance of returns. Then, a time-varying maximum drawdown is considered to adjust the risk aversion, which can efficiently cope with the limit loss problems. Both simulation and empirical results show that trading strategies based on the proposed method can provide better out-of-sample performance than the existing methods.

E0698: **Robust M-estimation for high dimensional regression on large panel data**
Presenter: **Huanjun Zhu**, Xiamen University, China

Robust estimation of high dimensional regression coefficient vectors for large panel data is considered. Unlike the high dimensional regression model, cross-sectional dependence makes a robust estimation for large panel data more complicated. To describe cross-sectional dependence, some cross-sectional dependence is allowed even after taking out observed common factors. To achieve robustness against heavy-tailed sampling distributions, a robust M-estimator by assuming regular conditions on general loss functions is considered. The asymptotic Bahadur representation is provided, and simultaneously asymptotic joint distribution for the high-dimensional regression coefficient vector and factor loading vector is established. Thorough numerical results on simulated and real datasets illustrate commonly used M-estimators that outperform the least-squares method on heavy-tailed distributed data.

E0772: **Non-axis aligned space partitioning**
Presenter: **Shufei Ge**, ShanghaiTech University, China

Space partitioning methods, such as decision trees and the Mondrian processes (MP), are often used to model relational data in multi-dimensional space. However, the flexibility of these methods is often limited by the requirement that the cuts be axis aligned. The binary space partitioning (BSP)-tree process was recently introduced as a generalization of the MP for space partitioning with non-axis aligned cuts in the two-dimensional space. While in practice, usually, there are more than two predictors. Motivated by the need for non-axis aligned cuts for multi-dimensional data, the MP was generalized in an arbitrary space. A sequential Monte Carlo algorithm for inference is derived, and random forest versions are provided. The proposed process is self-consistent, allowing oblique cuts and enabling complex inter-dimensional dependence to be captured in multi-dimensional space. In addition, a novel parallel Bayesian nonparametric approach was proposed to partition images with curves, enabling complex object shapes to be acquired.

EO212 Room 605 TOPICS IN STATISTICAL LEARNING Chair: Archer Yang

E0348: **Using multi-source data to estimate subgroup effects in an external or external target population**

Presenter: **Guanbo Wang**, Harvard University, United States

One major challenge in estimating effect heterogeneity is that the sample size of the data used is typically not enough to capture how effects vary according to the effect modifiers precisely. Therefore, there is interest in synthesizing evidence across multi-source data (e.g., multi-centre trials, meta-analyses of randomized trials, pooled analyses of observational cohorts) to improve the precision of estimators of heterogeneous treatment efficacy. Furthermore, when combining information from multi-source data, the samples typically do not represent a common target population of substantive interest. This raises the question of combining information from multi-source data in an interpretable way in the context of some meaningful target population of interest while using evidence across multi-source data to improve efficiency. Methods are developed and evaluated for using multi-source data to estimate subgroup treatment effects in an external target population or the populations underlying the data sources. A doubly robust estimator is proposed that, under mild conditions, is non-parametrically efficient and allows for nuisance functions to be estimated using machine learning methods. The methods are illustrated in meta-analyses of randomized trials for schizophrenia and bipolar disorder.

E0851: **Density-based clustering with kernel diffusion**

Presenter: **Chao Zheng**, University of Southampton, United Kingdom

Finding a suitable density function is essential for density-based clustering algorithms such as DBSCAN and DPC. A naive density corresponding to the indicator function of a unit d-dimensional Euclidean ball is commonly used in these algorithms. Such density suffers from capturing local features in complex datasets. To tackle this issue, a new kernel diffusion density function is proposed, which is adaptive to data of varying local distributional characteristics and smoothness. Furthermore, a surrogate is developed that can be efficiently computed in linear time and space and prove that it is asymptotically equivalent to the kernel diffusion density function. Extensive empirical experiments on benchmark and large-scale face image datasets show that the proposed approach not only achieves a significant improvement over classic density-based clustering algorithms but also outperforms the state-of-the-art face clustering methods by a large margin.

E1069: **Feature screening with conditional rank utility for big-data classification**

Presenter: **Chen Xu**, University of Ottawa, Canada

Feature screening is a commonly-used strategy to eliminate irrelevant features in high-dimensional classification. When one encounters big datasets with both high dimensionality and huge sample size, the conventional screening methods become computationally costly or even infeasible. A novel screening utility, Conditional Rank Utility (CRU), is introduced, and a distributed feature screening procedure for the big-data classification is proposed. The proposed CRU effectively quantifies the significance of a numerical feature on the categorical response. Since CRU is constructed based on the ratio of the mean conditional rank to the mean unconditional rank of a feature, it is robust against model misspecification and the presence of outliers. Structurally, CRU can be expressed as a simple function of a few component parameters, each of which can be distributively estimated using a natural unbiased estimator from the data segments. Under mild conditions, it is shown that the distributed estimator of CRU is fully efficient in terms of the probability convergence bound and the mean squared error rate; the corresponding distributed screening procedure enjoys the sure screening and ranking properties. Extensive numerical examples support the promising performances of CRU-based screening.

EO078 Room 606 RECENT DEVELOPMENTS IN COPULA MODELING Chair: Pavel Krupskiy

E0563: **Towards a universal representation of statistical dependence**

Presenter: **Gery Geenens**, University of New South Wales, Australia

Dependence is undoubtedly a central concept in statistics. Though, it proves difficult to locate in the literature a formal definition which goes beyond a self-evident interpretation of "dependence = non-independence", which quickly proves inadequate. For example, quantifying dependence between two variables appears essential in many situations; yet, if dependence is to be quantifiable, then the above non-independence definition falls short, and this is without any obvious substitute. This absence has allowed the term "dependence" and its declination to be used vaguely and indiscriminately for qualifying a variety of disparate notions, leading to numerous incongruities. Arguing that research on such a fundamental topic would benefit from a slightly more rigid framework, this work suggests a general definition of the dependence between two random variables defined on the same probability space. Natural enough for aligning with intuition, that definition is still sufficiently precise for allowing unequivocal identification of a "universal" representation of the dependence structure of any bivariate distribution regardless of its nature (discrete, continuous, mixed, hybrid). Links between this representation and familiar concepts are highlighted. The role of copulas will also be discussed from that perspective, showing that copulas provide a sensible approach for analysing and modelling dependence in a continuous vector but cannot be justified outside that framework.

E0639: **Conditional dependence models under covariate measurement error**

Presenter: **Elif Acar**, University of Manitoba, Canada

Co-authors: Kaiqiong Zhao

In many applications, covariates are subject to measurement error. While there is a vast literature on measurement error problems in regression settings, very little is known about the impact of covariate measurement error on the dependence parameter estimation in multivariate models. The latter problem is addressed using a conditional copula model, and it is shown that the dependence parameter estimates can be significantly biased if the covariate measurement error is ignored in the analysis. The underlying bias pattern from the direction and magnitude of marginal effect sizes is identified and an analytical bias correction method for the special case of the Gaussian copula is introduced. For general conditional copula models, a likelihood-based correction method is proposed, in which the likelihood function is computed via Monte Carlo integration. The consistency of the bias-corrected estimators is established. Numerical studies confirm that the proposed bias-correction methods achieve accurate estimation of the dependence parameter.

E1040: **On factor copula-based mixed regression models**

Presenter: **Pavel Krupskiy**, Melbourne University, Australia

Co-authors: Bouchra Nasri, Bruno N Remillard

A copula-based method for mixed regression models is introduced, 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 bivariate copula. It is shown how to estimate the parameters of the copula and the parameters of the margins, and the asymptotic behaviour of the estimation errors is found. Numerical experiments are performed to assess the precision of the estimators for finite samples. An example of an application is given using dengue data from several countries.

EO170 Room 701 RECENT DEVELOPMENTS IN DEGRADATION ANALYSIS AND RELATED TOPICS I Chair: Tsai-Hung Fan

E0250: **Optimum test planning for heterogeneous Wiener processes**
Presenter: **Ya-Shan Cheng**, National Tsing Hua University, Taiwan
Co-authors: Chien-Yu Peng

Degradation models based on heterogeneous Wiener processes are commonly used to assess the lifetime information for highly reliable products. In general, an optimum test plan under limited resources is found by numerical methods for the heterogeneous Wiener process. However, the numerical search for optimum test plans can not avoid the time-consuming and locally optimum issues. To overcome these difficulties, an explicit expression is derived for decision variables (such as the termination time, the number of measurements, and sample size) of the D- and V-optimum test plans with cost constraints. The theoretical results not only ensure the globally optimum test plan but also provide clear insights into decision variables affected by the model parameters and experimental costs. Some numerical examples are presented to support the efficiency and applicability of the optimum test plans.

E0665: **Estimating the useful life of lithium-ion battery pack based on a mixed effect degradation model**
Presenter: **Shuen-Lin Jeng**, National Cheng Kung Univeraity, Taiwan

Generally speaking, the reliability of a module composed of components can be directly calculated by the system reliability theory through the component life distribution, but this calculation cannot be directly applied to the actual conditions of lithium-ion battery modules. The failure time distribution of a single Li-ion cell is derived from the experimental discharge degradation path of a single Li-ion cell, and then the system reliability theory combined with a simplified electrical model in electrochemistry is used to develop a mixed effect capacitance degradation model for estimating the remaining life of a Li-ion battery pack. The proposed method can use less than 50% of the entire life cycle data of a single cell to determine the remaining life distribution of a lithium-ion battery cell. The reliability and lifetime of packs with different series-parallel configurations can then be calculated. The importance of consistency in the life distribution of lithium-ion single-cell batteries is verified, and the optimal series-parallel connection method for lithium-ion battery packs is compared.

E0847: **Local influence on gamma process and trend gamma process**
Presenter: **Yufen Huang**, National Cheng Kung University, Taiwan

The gamma process (GP) is widely used when the degradation path is strictly increasing. For some circumstances, the GP is not able to successfully describe the degradation path. Hence, random effects are considered in the GP model (REGP) for resolving this problem. Alternatively, a trend gamma process (TGP) has been previously proposed, which integrates the merits of the trend function into a GP model attempting to overcome this obstacle. Case diagnostics play an important role in statistical modelling. For example, during the model fitting process, suspicious observations can greatly influence modelling and forecasting results. Consequently, the detection of such aberrant observations becomes an essential task. To our knowledge, a study on influence analysis for GP, REGP and TGP models has not been explored in the literature. Local influence has been previously proposed to assess the local effect of small perturbations in regression models. Local influence on degradation paths in GP, REGP and TGP models are developed as tools for case diagnostics. Simulation studies and real data examples show the proposed method provides a good tool for outlier/influential path detection as well as an auxiliary diagnosis for assuring the quality of data while using GP, REGP and TGP models.

EO047 Room 702 STATISTICS IN NEUROSCIENCE Chair: Russell Shinohara

E0176: **Beyond ComBat: Next-generation harmonization methods for multi-centre neuroimaging studies**
Presenter: **Russell Shinohara**, University of Pennsylvania, United States

Magnetic resonance imaging (MRI) studies often involve large multi-centre designs. To address this, an increasingly commonly used approach is ComBat, which was first proposed in the genomics literature. Our group has found these methods helpful but sometimes insufficient in complex study designs, mainly when employing joint or predictive modelling. To address this, a broader framework has been developed for multi-scanner harmonization, allowing for flexible, non-linear, and multivariate modelling in various imaging science settings.

E0208: **Arguments for the biological and predictive relevance of the proportional recovery rule**
Presenter: **Jeff Goldsmith**, Columbia University, United States

The proportional recovery rule (PRR) posits that most stroke survivors can expect to reduce a fixed proportion of their motor impairment. The PRR explicitly relates change scores to baseline values as a statistical model. This approach arises in many scientific domains but can potentially introduce artefacts and flawed conclusions. Approaches are described that can assess associations between baseline and changes from baseline while avoiding artefacts due to mathematical coupling or regression to the mean. Methods that can compare different biological models of recovery are also described. Across several real datasets in stroke recovery, evidence for non-artifactual associations between baseline and change is found, and support for the PRR is compared to alternative models. A statistical perspective is also introduced that can be used to assess future models in the conclusion that the PRR remains a biologically relevant model of stroke recovery.

E0470: **Partition learning for functional neuro connectivity**
Presenter: **Emily Hector**, North Carolina State University, United States

Motivated by the need to model joint dependence between regions of interest in functional neuro connectivity for efficient inference, a new Bayesian clustering approach is proposed for correlation structures of high-dimensional Gaussian outcomes. The key technique is a Dirichlet process that clusters correlation sub-matrices into independent subgroups of outcomes, thereby naturally inducing sparsity in the whole brain connectivity matrix. A new split-merge algorithm is employed to improve the mixing of the sampling chain shown empirically to recover both uniform and true Dirichlet partitions with high accuracy. The approach's performance is investigated through extensive simulations. Finally, the proposed approach is used to group regions of interest into functionally independent sub-groups in the Autism Brain Imaging Data Exchange participants with autism spectrum disorder and attention-deficit/hyperactivity disorder.

EO158 Room 703 ADVANCES IN CONTEMPORARY SPATIAL AND SPATIOTEMPORAL DATA ANALYSIS Chair: Shan Yu

E1230: **Estimation and inference of quantile spatially varying coefficient models over complicated domains**
Presenter: **Myungjin Kim**, Kyungpook National University, Korea, South
Co-authors: Lily Wang, Huixia Judy Wang

Spatially varying coefficient models provide a flexible extension of linear regression models to capture the non-stationarity of regression coefficients across space. However, its research mostly focuses on mean regression settings. Quantile regression gives a more comprehensive analysis of the effect of the predictors on the response when one is interested in the full distributional properties of the response or when assumptions on the mean regression are violated. A flexible quantile spatially varying coefficient model is introduced to assess how conditional quantiles of the response depend on covariates, allowing the coefficient function to vary with spatial location. The model can be used to explore spatial non-stationarity of regression relationships for heterogeneous spatial data distributed over a domain of a complex or irregular shape. To estimate the coefficient functions of the model over a complex spatial domain, a quantile regression method is proposed that adopts the bivariate penalized spline technique to approximate the unknown functional coefficients. The L2 convergence of the proposed estimator with an optimal convergence rate under some regularity conditions is established. Also, an efficient algorithm based on the alternating direction method of multipliers is developed to solve the optimization problem. Numerical studies examine the finite sample performance of the proposed method.

E1294: **Identifying localized dynamic changes in large-scale spatio-temporal data through generalized nonparametric regression**
Presenter: **Shan Yu**, University of Virginia, United States
Co-authors: Yuda Shao

The rapid advancement of modern technology has led to the generation of vast amounts of large-scale spatial-temporal datasets, which offer valuable insights into human behaviour. The task of identifying when and where changes occur in spatial-temporal processes has gained significant attention in recent years. A generalized spatial-temporal modelling framework is proposed that captures the trends in the spatiotemporal process and identifies regions with quick changes. To achieve this, tensor product splines over triangular prismatic partitions are utilized to approximate the unknown spatial-temporal trend. A piecewise-penalty function is then imposed to efficiently identify the regions with dynamic changes, employing a computationally efficient algorithm. Simulation studies are conducted, and the proposed method is applied to mobile location data in Baltimore, demonstrating its effectiveness in efficiently recovering regions with dynamic changes over time.

E1311: **Online change point detection in high-dimensional vector auto-regressive models**
Presenter: **Abolfazl Safikhani**, University of Florida, United States

Online change point detection consists of sequentially monitoring a time series and raising the alarm if a shift in the data distribution is detected. Given data generated by a high-dimensional vector auto-regressive model, an algorithm is proposed to detect changes in the model transition matrices in an online format. The algorithm consists of two main steps. First, the estimation of transition matrices and variance of error terms are calculated by applying regularization methods to the training data. As new batches of data are observed, a specific test statistic is calculated in the second step to check whether the transition matrices have changed. Asymptotic normality of the test statistic in the regime of no change points is established under mild conditions. The alarm will be raised if the test statistic is larger than a certain quantile of standard normal distribution. Further, the relationship between the power of the test and jump size is established, and it is verified that for a large enough jump size, the power of the test converges to one. The proposed algorithm is memory-saving since it only requires storing the estimations and new batches of data in computer memoralgorithm's effectivenessgorithm is confirmed empirically through various simulation settings and comparisons with some competing methods. Finally, applications to analyze shocks in S&P 500 data and to detect the timing of seizures in EEG data are discussed.

EO081 Room 704 DESIGN AND ANALYSIS OF COMPLEX EXPERIMENTS: THEORY AND APPLICATIONS Chair: MingHung Kao

E0315: **Indicator-based Bayesian variable selection for Gaussian process models in computer experiments**
Presenter: **Ray-Bing Chen**, National Cheng Kung University, Taiwan
Co-authors: Fan Zhang, Ying Hung, Xinwei Deng

Gaussian process (GP) models are commonly used to analyse computer experiments. Variable selection in GP models is of significant scientific interest, but existing solutions remain unsatisfactory. For each variable in a GP model, there are two potential effects with different implications: one is on the mean function, and the other is on the covariance function. However, most research on variable selection for GP models has focused only on one of the effects. To tackle this problem, an indicator-based Bayesian variable selection procedure is proposed to consider the effects of both the mean and covariance functions. A variable is defined as inactive if both effects are not significant, and an indicator is used to represent whether the variable is active or not. The proposed method adopts different prior assumptions for active variables to capture the two effects. Both simulations and real applications in computer experiments evaluate the performance of the proposed method.

E0750: **Optimal designs for sparse functional data**
Presenter: **MingHung Kao**, Arizona State University, United States

Sparse functional data analysis (FDA) is powerful for making inferences on the underlying random function when noisy observations are collected at sparse time points. Knowledge of optimal designs that allow the experimenters to collect informative, functional data is crucial for a precise inference. We propose a framework for selecting optimal designs to precisely predict functional principal and empirical component scores. A relevant generalization of previous results on the design for predicting individual response curves is given. Optimal designs are obtained, and evaluate the performance of commonly used designs. It is demonstrated that without a judiciously selected design, statistical efficiency can be lost.

E0900: **A systematic design construction and analysis for cost-efficient order-of-addition experiment**
Presenter: **Frederick Kin Hing Phoa**, Academia Sinica, Taiwan
Co-authors: Jing-Wen Huang

A systematic design construction method is proposed for cost-efficient order-of-addition (OofA) experiments and its corresponding statistical models for analyzing experimental results. Specifically, these designs consider the effects of two successive treatments. Each pair of level settings from two different factors in our design matrix appears exactly once to achieve cost-efficiency. Compared to designs in recent studies of OofA experiments, this design can conduct experiments of one or more factors, so practitioners can insert a placebo or choose different doses as level settings when this design is used as their experimental plans. An experimental analysis based on this design performs better than those based on the minimal-point design and Bayesian D-optimal design with pairwise-order modelling in identifying the optimal order.

EO155 Room 705 RECENT ADVANCES IN STATISTICAL METHODS AND THEORY Chair: Mengyu Xu

E0597: **Joint estimation of precision matrices for high-dimensional time series with long-memory**
Presenter: **Jongik Chung**, University of Central Florida, United States
Co-authors: Qihu Zhang, Cheolwoo Park

The focus is on the simultaneous estimation of multiple precision matrices for high-dimensional time series with long memory. The motivating example is a time series of resting-state functional magnetic resonance imaging (fMRI) data collected from multiple subjects. Estimating the brain network for each subject and a common structure representative of a group of subjects is of interest. A few approaches are considered for simultaneously estimating individual and group precision matrices for long-memory time series data using weighted aggregation. The convergence rates of the precision matrix estimators for various norms and their expectations under a sub-Gaussian or heavy-tailed assumption are examined. The empirical performance is demonstrated via simulated examples and resting-state fMRI data.

E0674: **Robust sufficient dimension reduction and sufficient variable selection via distance covariance**
Presenter: **Teng Zhang**, University of Central Florida, United States
Co-authors: Hsin-Hsiung Huang

Sufficient dimension reduction (SDR) using distance covariance (DCOV) was recently proposed as an approach to dimension-reduction problems. Compared with other SDR methods, it is model-free without estimating link function and does not require any particular distributions on predictors.

However, the DCOV-based SDR method optimises a non-smooth and nonconvex objective function over the Stiefel manifold. To tackle the numerical challenge, the original objective function is equivalently formulated into a difference of convex functions program and develop an efficient algorithm based on the projection on the Stiefel manifold. In addition, the algorithm can also be readily extended to sufficient variable selection using distance covariance. Finally, the convergence property of the proposed algorithm under some regularity conditions is established. Simulation and real data analysis show the algorithm drastically improves the computation efficiency and is robust across various settings compared with the existing methods.

E0769: **Online change point detection in high-dimensional factor models**
Presenter: **Mengyu Xu**, University of Central Florida, United States
Co-authors: Mahdi Mirhosseini

The focus is monitoring high-dimensional data for potential change points in the mean vector. The data is assumed to be generated from a factor model. Change point detection is studied on the fixed dimensional factors and sparse idiosyncratic noises respectively. A Gaussian approximation result for the high-dimensional noise process is developed. The method is applied to an economic dataset for change point detection.

EO062 Room 708 RECENT ADVANCE IN NEUROIMAGING STUDIES Chair: Yi Zhao

E0534: **Bayesian image mediation analysis**
Presenter: **Jian Kang**, University of Michigan, United States

Neuroimaging data presents unique challenges for mediation analysis due to its high dimensionality, complex spatial correlations, sparse activation patterns, and relatively low signal-to-noise ratio. A new Bayesian image mediation analysis (BIMA) method that employs a spatially varying coefficient structural equation model is proposed to address these challenges. A soft-thresholded Gaussian process (STGP) is used for prior specifications of the spatially varying coefficients, which enables large prior support for sparse and piece-wise smooth functions. The spatially varying mediation effects of the exposure on the outcome mediated through imaging mediators under the potential outcome framework are defined. Posterior consistency is established for spatially varying mediation effects and selection consistency on important regions that contribute to the mediation effects. An efficient posterior computation algorithm for BIMA is developed and scalable for large-scale imaging data analysis. As demonstrated through simulations, BIMA improves estimation accuracy and computational efficiency for high-dimensional mediation analysis compared to existing methods. Additionally, BIMA is applied to analyze behavioural and fMRI data in the Adolescent Brain Cognitive Development (ABCD) study and infer the mediation effects of parental education level on children's general cognitive ability mediated through working memory brain activities.

E0657: **Measuring cross-channel information transfer in the frequency domain through spectral transfer entropy**
Presenter: **Paolo Victor Redondo**, King Abdullah University of Science and Technology, Saudi Arabia
Co-authors: Raphael Huser, Hernando Ombao

Brain connectivity reflects how different regions of the brain interact during the performance of a cognitive task. Studying brain signals, such as electroencephalograms (EEG), may be explored via transfer entropy (TE). This information-theoretic causal measure covers any form of relationship (beyond linear) between variables. To improve the utility of TE, a novel methodology is proposed to capture cross-channel information transfer in the frequency domain. A new causal measure, spectral transfer entropy (STE), is introduced to quantify the magnitude and direction of information flow from a frequency-band oscillation of a channel to an oscillation of another channel. In contrast with previous works on TE in the frequency domain, this method is differentiated by considering an extreme value perspective that employs the maximum magnitude of filtered series within time blocks. The main advantages of this approach are that it is robust to the issues of linear filtering and allows adjustments for multiple comparisons to control family-wise error rates. Another novel contribution is a simple yet efficient estimation method based on the combination of vine copulas and extreme value theory that enables estimates to capture zero (boundary point) is illustrated without the need for bias adjustments. Lastly, the advantage of our measure through numerical experiments and interesting and novel findings are provided in the analysis of EEG recordings linked to a visual task.

E0866: **Statistical challenges for large neuroimaging cohort studies**
Presenter: **Haochang Shou**, University of Pennsylvania, United States

With the increasing need for big data analytics in medical imaging, integrating data from multiple studies and various biological domains has become critical to better understanding complex human diseases. Analysis of such data is challenging due to the existence of site differences that could mask the biological associations of interest. The different data property of various modalities further poses difficulties in conducting integrative analyses to evaluate the joint relationship among modalities. Several most recent developments in statistical harmonization methods in large neuroimaging studies under various data modalities will be discussed. These approaches are designed to mitigate site differences that exist in mean, variance, and covariance structures in structural and functional imaging outcomes. A novel distance-based regression model, referred to as Similarity-based Multimodal Regression (SiMMR), will be discussed, enabling simultaneous regression of multiple modalities through their distance profiles. The proposed method can detect associations of differing properties and dimensionalities in multimodal data, even with modest sample sizes. The methods are motivated by the iSTAGING (Imaging-based coordinate SysTem for AGing and NeurodeGenerative diseases) consortium, which brings together multisite neuroimaging studies to understand the complex and heterogeneous process of normal ageing and AD pathology.

EO215 Room 709 CURRENT DEVELOPMENTS IN INDUSTRIAL AND APPLIED STATISTICS Chair: Tsung-Jen Shen

E0466: **Multi-output Gaussian process based on neural kernel learning and its prediction applications**
Presenter: **Hsiang-Ling Hsu**, National University of Kaohsiung, Taiwan

The basic concepts of multivariate normal distribution, kernels, non-parametric models and conditional probability are essential to build a Gaussian process. The flexibility and uncertainty measure inherent in predictions makes the Gaussian process regression models widely applicable to various fields. However, Gaussian process regression is based on the similarity of the sample function space and independently builds each predictive model but ignores the correlation between time points. On the other hand, the prediction results of the Gaussian process are influenced by the structure assumptions of the covariance function. Hence, a multi-output Gaussian process regression model attempts are proposed to combine different kernel functions through a neural kernel network framework to acquire the complicated correlation patterns between data for predictions. The electric load data of Kyushu in Japan and the apparent temperature of Lingya, Kaohsiung City, Taiwan, are analyzed based on the proposed method with two kinds of evaluation indexes, the root mean square error (RMSE) and mean absolute percentage error (MAPE), to measure the prediction performances. Compared with persistence, seasonal autoregressive integrated moving average (SARIMA) and traditional Gaussian process regression (GPR), the experimental results show that the proposed method possesses low MAPE and RMSE in forecasting the electricity load and apparent temperature 2020.

E0719: **Design construction and model selection for small mixture-process variable experiments with high-dimensional model terms**
Presenter: **Chang-Yun Lin**, National Chung Hsing University, Taiwan
Co-authors: Kashinath Chatterjee

The design construction and model selection for mixture-process variable experiments where the number of variables is large is considered. For such experiments, the generalized least squares estimates cannot be obtained, and hence it will be difficult to identify the important model terms. To overcome these problems, the generalized Bayesian-D criterion is employed to choose the optimal design and the Bayesian analysis method is applied to select the best model. Two algorithms are developed to implement the proposed methods. A fish-patty experiment demonstrates how the Bayesian approach can be applied to a real experiment. Simulation studies show that the proposed method has a high power to identify important terms and well control the type I error.

E0551: **A Bayesian approach to process optimization on data with multi-stratum structure**
Presenter: **Po Yang**, University of Manitoba, Canada

Multistratum design arises naturally in industrial experiments due to the inconvenient and impractical complete randomization. Most research has concentrated on finding optimal multi-stratum designs with high parameter estimation efficiencies. Accounting for the model uncertainty, the Bayesian model averaging method and predictive approach are applied to investigate the optimization problem for data with a multi-stratum structure. With the posterior probabilities of models as weights, the weighted average of the predictive densities of the response overall potential models are considered. The goal of the optimization is to identify the values of the factors that result in a maximum probability of a response in a given range. The method is illustrated with two examples.

Tuesday 01.08.2023 10:10 - 11:50 Parallel Session B – EcoSta2023

EO183 Room 02 ADVANCES OF HIGH-DIMENSIONAL STATISTICS IN BIOLOGICAL AND BIOMEDICAL RESEARCH Chair: Zhao Ren

E0790: **Integrative structural learning of mixed graphical models via pseudo-likelihood**
Presenter: **Yuping Zhang**, University of Connecticut, United States
Co-authors: Qingyang Liu

Markov Random Field is a common tool to characterize interactions among a fixed collection of variables. In recent biomedical research, new concerns have arisen about the discovery of regulatory and co-expression relationships among different types of features across multiple biological classes. Consequently, a data integration framework is proposed to learn multiple mixed graphical models simultaneously and jointly. To address the common asymmetry problem in neighbourhood selection, a new estimator is constructed using regularized pseudo-likelihood, which produces sym- metric and consistent estimates of network topologies. The practical merits of the method are demonstrated through learning synthetic networks and constructing gene regulatory networks from TCGA data.

E0957: **A data adaptive nonparametric procedure to define and assess reproducibility across high-throughput studies**
Presenter: **Austin Ellingworth**, Colorado State University, United States
Co-authors: Wen Zhou, Debashis Ghosh, Zhigen Zhao

Reproducibility is a fundamental aspect of research that ensures the validity of findings. Although a consensus on assessing reproducibility remains elusive, in high-throughput studies, reproducibility is often defined as the consistency of test results across experiments. Most existing approaches either rely on stringent parametric assumptions of summary statistics or only focus on the hypothesis-wise alignment of summary statistics but overlook the experiment-wise heterogeneity. Inspired by previous work, a function based on the ranks of summary statistics from each experiment is introduced to define a notion of reproducibility and also to identify reproducible discoveries. The proposed nonparametric procedure takes into account both the signal strength and experiment-wise heterogeneity. Examining the geometry of the space of ranks of summary statistics and utilizing the blessing negative association of ranks, we introduce a novel procedure for identifying reproducible findings while controlling the false discovery rate (FDR). Our method surpasses existing approaches in terms of power, effectively controlling FDR under relatively mild assumptions. We validate our theoretical findings through extensive simulations and apply our approach to two large-scale TWAS datasets, uncovering reproducible features. Overall, our innovative method significantly advances reproducibility research and offers valuable practical implications.

E0487: **Heteroskedastic sparse PCA in high dimensions**
Presenter: **Zhao Ren**, University of Pittsburgh, United States

Principal component analysis (PCA) is one of the most commonly used dimension reduction and feature extraction techniques. Though it has been well-studied for high-dimensional sparse PCA, little is known when the noise is heteroscedastic, ubiquitous in many scenarios, like biological sequencing data and information network data. An iterative algorithm is proposed for sparse PCA in the presence of heteroskedastic noise, which alternatively updates the estimates of the sparse eigenvectors using the power method with adaptive thresholding in one step. In the other step, it imputes the diagonal values of the sample covariance matrix to reduce the estimation bias due to heteroskedasticity. Our procedure is computationally fast and optimal under the generalized spiked covariance model, assuming the leading eigenvectors are sparse. A comprehensive simulation study demonstrates its robustness and effectiveness in various settings.

E0813: **Supervised capacity preserving mapping: A clustering guided visualization method for scRNAseq data**
Presenter: **Yuying Xie**, Michigan State University, United States

The rapid development of scRNA-seq technologies enables the exploration of the transcriptome at the cell level on a large scale. However, current visualization methods, including t-SNE and UMAP, are challenged by the limited accuracy of rendering the geometric relationship of populations with distinct functional states. Most visualization methods are unsupervised, leaving out information from the clustering results or given labels. This leads to the inaccurate depiction of the distances between the bona fide functional states. In particular, UMAP and t-SNE are not optimal for preserving the global geometric structure. They may result in a contradiction that clusters with near distance in the embedded dimensions are, in fact, further away in the original dimensions. Besides, UMAP and t-SNE cannot track the variance of clusters. Through the embedding of t-SNE and UMAP, the variance of a cluster is not only associated with the true variance but also is proportional to the sample size. SupCPM, a robust supervised visualization method, is presented, which separates different clusters, preserves the global structure and tracks the cluster variance. Compared to six visualization methods using various datasets, supCPM shows improved performance than other methods in preserving the global geometric structure and data variance. Overall, supCPM provides an enhanced visualization pipeline to assist the interpretation of functional transition and accurately depict population segregation.

EO059 Room 03 MODERN STATISTICAL METHODS FOR DATA ANALYSIS Chair: Kuo-Jung Lee

E0471: **Marginal quantile regression for longitudinal data with time-dependent covariates and its applications**
Presenter: **I-Chen Chen**, US Centers for Disease Control and Prevention, United States

Modelling the within-subject correlation structure using marginal quantile regression for longitudinal data can be difficult unless a working independence structure is utilized. Although this approach guarantees consistent estimators of the regression coefficients, it may lead to less efficient regression parameter estimation when data are highly correlated. Therefore, several marginal quantile regression methods have been proposed to improve regression parameter estimation. In a longitudinal study, some of the covariates may change their values over time, and the topic of time-dependent covariates has not been explored in the literature. Therefore, an approach for marginal quantile regression in the presence of time-dependent covariates, which includes a strategy, is proposed to select a working type of time dependency. A simulation study demonstrates that the proposed method can potentially improve power relative to the independence estimating equations approach due to the reduction of mean squared error. The proposed method in a working example and its extension to exposure and biomonitoring data from occupational studies is also discussed.

E0501: **Linear measurement models with replications in independent variable**
Presenter: **Jia-Ren Tsai**, Fu Jen Catholic University, Taiwan
Co-authors: Lei Ting-I

The effect of measurement errors increases with uncertainty in parameter estimates. The focus is investigating independent variables with replicated observations in the linear heteroscedastic measurement errors model. The adjusted least squares method can be extended, and the asymptotic covariance matrix of parameters also be constructed by using the sandwich method and the model-based expectation method based on unbiased estimating equations theory. A confidence interval is also constructed for a single slope parameter; and the joint parameters of the intercept and parameter terms. A Monte Carlo simulation study has been performed to assess the relationship and influence of the average reliability ratio, sample size and replicates under the variation of heterogeneous measurement errors. The application of the proposed method is also illustrated with a real dataset.

E0973: **Bayesian structure selection approaches for multiple binary responses via multi-task learning**
Presenter: **Chi-Hsiang Chu**, Tunghai University, Taiwan

Bayesian structure selection problems for the categorical response are addressed. It is focused on solving the selection problem for multiple binary responses, and the probit model for each response is used. Here, the group structure is considered with sparsity property on the rows of the coefficient matrix where each row corresponds to one variable. Then the relevant variables for the responses are identified, and the selection problems can be treated as the multi-task learning problem. The effectiveness of the proposed method will be demonstrated through simulation studies.

E1161: **Fast Bayesian whole brain connectivity estimation by GPU-enhanced Gaussian processes**
Presenter: **Hakmook Kang**, Vanderbilt University, United States
Co-authors: Yuting Mei, Ilwoo Lyu, Kim Albert, Brian Boyd, Bennett Landman, Warren Taylor

A previous Bayesian spatiotemporal model approach has been expanded to significantly reduce the computation burden by employing GPU (Graphics Processing Unit) computing and Gaussian Process to model the intra-voxel spatial correlation in each ROI (region of interest). A Bayesian double-fusion technique was 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 us to precisely investigate the relationship between FC and SC or alterations in white matter microstructural integrity. The method is applied to multi-subject data ($n = 45$) with depression ($n = 20$) and without depression ($n = 25$) to examine how FC differences are related to cognitive task performance in depression.

EO180 Room 04 RECENT ADVANCES IN THE ANALYSIS OF CENSORED DATA Chair: Feng-Chang Lin

E0542: **Deep neural network based accelerated failure time models with high-dimensional data**
Presenter: **Gwangsu Kim**, Jeonbuk National University, Korea, South

An accelerated failure time (AFT) model assumes a log-linear relationship between failure times and a set of covariates. In contrast to other popular survival models that work on hazard functions, the effects of covariates are directly on failure times, whose interpretation is intuitive. Also, deep neural networks (DNN) have received focal attention over the past decades and have achieved remarkable success in various fields. DNNs have a number of notable advantages and have been shown to be particularly useful in addressing nonlinearity. By taking advantage of this, the method proposes to apply DNN in fitting AFT models using a Gehan-type loss combined with a sub-sampling technique. An extensive simulation study investigates the finite sample properties of the proposed DNN and rank-based AFT model (DeepR-AFT). DeepR-AFT shows superior and interesting performance, especially in high-dimensional data.

E0567: **Response-adaptive randomization for recurrent and terminal events data with a composite endpoint**
Presenter: **Pei-Fang Su**, National Cheng Kung University, Taiwan

Recurrent event and terminal event data commonly arise in clinical and observational studies. A composite endpoint has been used as a possible assessment to evaluate the efficacy of a treatment effect for both types of events, particularly when faced with high costs and a more extended follow-up study. To model recurrent event processes complicated by the existence of a terminal event, joint frailty modelling has been typically employed. The objective was to develop target-driven response adaptive randomization strategies using a composite endpoint based on joint frailty modelling. First, a balanced, randomized design was implemented, and then the adaptive response randomization was investigated. The former is intuitively adopted first, while the latter is expected to be desirable and ethical in allocating more subjects to the more effective treatment. The results show that the proposed procedures using a composite endpoint are capable of reducing the number of trial participants who receive inferior treatment while simultaneously reaching a desired optimal target as compared to a balanced, randomized design. The R shiny application for calculating the sample size and allocation probabilities is also available.

E0984: **Improved semiparametric estimation of the proportional rate model with recurrent event data**
Presenter: **Ming-Yueh Huang**, Academia Sinica, Taiwan
Co-authors: ChiungYu Huang

The pseudo-partial likelihood method, known for its robustness, marginal interpretations, and ease of implementation, has become the default method for analyzing recurrent event data using Cox-type proportional rate models, as introduced in previous seminal papers. However, the pseudo-partial score function's construction does not account for dependency among recurrent events, leading to potential inefficiency. The asymptotic efficiency of weighted pseudo-partial likelihood estimation is explored, demonstrating that the optimal weight function depends on the unknown variance-covariance process of the recurrent event process and may lack a closed-form expression. Therefore, combining a set of pre-specified weighted pseudo-partial score equations is proposed using the generalized method of moments and empirical likelihood estimation rather than determining optimal weights. The findings indicate that significant efficiency improvements can be readily achieved without introducing additional model assumptions. Furthermore, the proposed estimation methods can be executed using existing software. Both theoretical and numerical analyses reveal that the empirical likelihood estimator is more desirable than the generalized method of moments estimator when the sample size is sufficiently large.

E1217: **Variable selection and estimation for the average treatment effect with error-prone confounders**
Presenter: **Li-Pang Chen**, National Chengchi University, Taiwan
Co-authors: Grace Yi

In the causal inference framework, the inverse-probability-weighting estimation method and its variants have been commonly employed to estimate the average treatment effect. Such methods, however, are challenged by the presence of irrelevant pre-treatment variables and measurement errors. Ignoring these features and naively applying the usual inverse probability-weighting estimation procedures may typically yield biased inference results. An inference method is developed for estimating the average treatment effect with those features considered. Theoretical properties are established for the resulting estimator, and numerical studies are carried out to assess the finite sample performance of the proposed estimator.

EO012 Room Virtual R01 SPATIAL STATISTICS Chair: Kapil Gupta

E0517: **Efficient divide-and-conquer approach for spatio-temporal modeling of real estate data**
Presenter: **Kapil Gupta**, Indian Institute of Management, Bangalore, India
Co-authors: Soudeep Deb

Statistical research in real estate markets has recently garnered attention from the perspective of understanding the spatiotemporal dynamics of house prices. Markov chain Monte Carlo (MCMC) is generally used for Bayesian inference in spatio-temporal modelling. However, standard techniques of MCMC are usually slow for large datasets such as real estate data due to the requirement of multiple passes through the entire data in each iteration. A divide-and-conquer spatiotemporal modelling approach is proposed to tackle this problem. The method involves partitioning the data into multiple subsets of sufficient locations and utilizing an appropriate Gaussian process model for each subset in parallel. The parameters corresponding to each subset are then combined to obtain the global parameters for the original problem. The proposed methodology allows us to assess the spatially varying impact of various factors on the house-price dynamics. It is also proved to be much faster than a conventional Bayesian approach. As a real life application of the proposed model, we analyze house price data from London from January 2011 to October 2019, covering 53 bi-monthly time points and 906 middle layer super output areas (MSOAs). The results furnish insightful analysis and render good predictive accuracy, as demonstrated by a cross-validation study.

E0518: **Bayesian multi-modal data integration**
Presenter: **Rajarshi Guhaniyogi**, Texas A & M university, United States

The focus is on a multi-modal imaging data application where structural/spatial information from grey matter (GM) and brain connectivity information in the form of a brain connectome network from functional magnetic resonance imaging (fMRI) are available for a number of subjects with different degrees of a neurodegenerative disorder (ND). The clinical/scientific goal becomes identifying brain regions of interest significantly related to the speech rate measure to gain insight into an ND pathway. Viewing the brain connectome network and GM images as objects, a flexible joint object response regression framework of network and GM images on the ND measure is developed. A novel joint prior formulation is proposed on the network and structural image coefficients to exploit network information of the brain connectome while leveraging the topological linkages among the connectome network and anatomical information from GM to draw inferences on brain regions significantly related to the ND. The principled Bayesian framework allows precise uncertainty characterisation in ascertaining a region being actively related to ND. Strategies to draw scalable Bayesian inference in these models will be discussed.

E0519: **Estimating atmospheric motion winds from satellite image data using space-time drift models**
Presenter: **Indranil Sahoo**, Virginia Commonwealth University, United States

Geostationary weather satellites collect high-resolution data comprising a series of images. The Derived Motion Winds (DMW) Algorithm is commonly used to process these data and estimate atmospheric winds by tracking image features. However, the wind estimates from the DMW Algorithm are often missing and do not come with uncertainty measures. Also, the DMW Algorithm estimates can only be half-integers since the algorithm requires the original and shifted data to be at the same locations to calculate the displacement vector between them. This motivates us to statistically model wind motions as a spatial process drifting in time. Using a covariance function that depends on spatial and temporal lags and a drift parameter to capture the wind speed and wind direction, the parameters are estimated by local maximum likelihood. The method allows us to compute standard errors of the local estimates, enabling spatial smoothing of the estimates using a Gaussian kernel weighted by the inverses of the estimated variances. Extensive simulation studies are conducted to determine the situations where our method performs well. The proposed method is applied to the GOES-15 brightness temperature data over Colorado and reduces the prediction error of brightness temperature compared to the DMW Algorithm.

E0520: **Matrix-free conditional simulations of Gaussian fields on a regular lattice**
Presenter: **Somak Dutta**, Iowa State University, United States
Co-authors: Debashis Mondal

Conditional simulations of random fields given observations at finitely many sampled locations provide realistic depictions of their local variations. However, traditional methods using dense matrix computations are not scalable because their storage and computational complexity scale at least quadratically with the sample size. A new matrix-free algorithm for conditional simulation for a large class of spatial mixed models will be discussed, and its use in spatial mapping and downscaling will be demonstrated.

EO187 Room Virtual R02 STATISTICAL METHODS IN HEALTH RESEARCH Chair: Jeong Hoon Jang

E0494: **Proformer: A hybrid macaron transformer model predicts expression values from promoter sequences**
Presenter: **Il-Youp Kwak**, Chung-Ang University, Korea, South

The breakthrough high-throughput measurement of the cis-regulatory activity of millions of randomly generated promoters provides an unprecedented opportunity to decode the cis-regulatory logic that determines the expression values systematically. An end-to-end transformer encoder architecture named Performer is developed to predict the expression values from DNA sequences. Performer used a Macaron-like Transformer encoder architecture, where two half-step feed-forward (FFN) layers were placed at the beginning and the end of each encoder block, and a separable 1D convolution layer was inserted after the first FFN layer and in front of the multi-head attention layer. The sliding k-mers from one-hot encoded sequences were mapped onto a continuous embedding, combined with the learned positional embedding and strand embedding (forward strand vs reverse complemented strand) as the sequence input. Moreover, Proformer introduced multiple expression heads with mask filling to prevent the transformer models from collapsing when training on a relatively small amount of data. It is empirically determined that this design had significantly better performance than the conventional design, such as using the global pooling layer as the output layer for the regression task. These analyses support the notion that Performer provides a novel learning method and enhances our understanding of how cis-regulatory sequences determine the expression values.

E0548: **Federated statistical learning with differential privacy**
Presenter: **Changgee Chang**, Indiana University, United States
Co-authors: Jaemu Heo, Jeonghun Kang, Taehwan Kim

While electronic health records (EHRs) offer great promises for advancing precision medicine, they suffer significant analytical challenges, which include the fact that it is often the case that patient-level data in EHRs cannot be shared across different institutions due to government regulations and/or institutional policies. A novel communication-efficient and privacy-preserving federated learning method is proposed to efficiently aggregate information from multiple datasets without exchanging individual patient-level data. Our new module INFEMBLER is presented, meaning an information assembler, which can extract and combine information carried in the perturbed MLE estimates from each remote database. The proposed approach allows proper statistical analyses from the linear model to various general models without sharing the raw patient-level datas. The proposed method's differential privacy properties and theoretical properties are investigated, and its performance is evaluated via simulations and real data analyses compared with several recently developed methods in the distributed statistical learning literature.

E0574: **A unified mediation analysis framework for integrative cancer proteogenomics with clinical outcomes**
Presenter: **Min Jin Ha**, Yonsei University, Korea, South

Multilevel molecular profiling of tumours and integrative analysis with clinical outcomes have enabled a deeper characterization of cancer treatment. Mediation analysis has emerged as a promising statistical tool to identify and quantify the intermediate mechanisms by which a gene affects an outcome. However, existing methods lack a unified approach to handle various types of outcome variables, making them unsuitable for high-throughput molecular profiling data with highly interconnected variables. A general mediation analysis framework is developed for proteogenomic data that includes multiple exposures and multivariate mediators on various effects scales as appropriate for continuous, binary and survival outcomes. The estimation method avoids imposing constraints on model parameters, such as the rare disease assumption while accommodating multiple exposures and high-dimensional mediators. Using kidney renal clear cell carcinoma proteogenomic data, it is identified genes that are mediated by proteins and the underlying mechanisms on various survival outcomes that capture short- and long-term disease-specific clinical characteristics.

E0594: **Assessing intra- and inter-method agreement of functional data**
Presenter: **Jeong Hoon Jang**, Yonsei University, Korea, South

Modern medical devices increasingly produce functional data whose sampling unit is a smooth continuous function defined over a time/spatial domain. A series of intraclass correlation coefficient (ICC) and concordance correlation coefficient (CCC) indices that can evaluate the reliability and reproducibility of medical devices producing functional data are proposed. Specifically, two versions of ICC and CCC indices are introduced. The first version consists of time-dependent ICC and CCC indices that can quantify the degrees of intra-method, inter-method and total (intra+inter) agreement that vary smoothly over time. The second version denotes their global counterparts, summarising agreement over the entire dime domain using a single measure. The proposed indices are formulated based on a multivariate multilevel functional model that represents indices in terms of truncated multivariate Karhunen-Loeve expansions, whose terms can be smoothly estimated by functional principal component analysis. Extensive simulation studies are performed to assess the finite-sample properties of the estimators. The proposed method is applied to Emory renal study data to evaluate the reliability and reproducibility of renogram curve data produced by a high-tech radionuclide image scan used to detect kidney obstruction non-invasively.

EO114 Room 102 NETWORK, GRAPHICAL MODEL AND MIXTURE MODELS Chair: Wenlin Dai

E0292: **Estimation and order selection for multivariate exponential power mixture models**
Presenter: **Zhenghui Feng**, Harbin Institute of Technology, Shenzhen, China
Co-authors: Xiao Chen, Heng Peng

Finite mixture models are promising statistical models for investigating the heterogeneity of a population. Using multivariate exponential power mixture models is considered for multivariate non-Gaussian density estimation and approximation. The penalized-likelihood method with a generalized EM algorithm to estimate locations, scale matrices, shape parameters, and mixing probabilities is proposed. Order selection is achieved simultaneously. Properties of the estimated order have been derived. Although it is mainly focused on the unconstrained scale matrix type in multivariate exponential power mixture models, three more parsimonious types of scale matrix have also been considered. Based on simulation and real data analysis, the performance implies the parsimony of the exponential power mixture models and verifies the consistency of order selection.

E0395: **A joint estimation approach to sparse additive ordinary differential equations**
Presenter: **Nan Zhang**, Fudan University, China
Co-authors: Muye Nanshan, Jiguo Cao

Ordinary differential equations (ODEs) are widely used to characterize the dynamics of complex systems in real applications. A novel joint estimation approach is proposed for generalized sparse additive ODEs where observations are allowed to be non-Gaussian. The new method is unified with existing collocation methods by simultaneously considering the likelihood, ODE fidelity and sparse regularization. A block coordinate descent algorithm is designed for optimizing the non-convex and non-differentiable objective function. The global convergence of the algorithm is established. The simulation study and two applications demonstrate the superior performance of the proposed method in estimation and improved performance of identifying the sparse structure.

E0403: **Combining smoothing spline with conditional Gaussian graphical model for density and graph estimation**
Presenter: **Yuedong Wang**, University of California - Santa Barbara, United States

Multivariate density estimation and graphical models play important roles in statistical learning. The estimated density can be used to construct a graphical model that reveals conditional relationships, whereas a graphical structure can be used to build models for density estimation. The goal is to construct a consolidated framework that can perform both density and graph estimation. Denote Z as the random vector of interest with density function $f(z)$. Splitting Z into two parts, $Z=(X,Y)$ and writing $f(z)=f(x)f(y|x)$, where $f(x)$ is the density function of X and $f(y|x)$ is the conditional density of $Y|X=x$. A semiparametric framework is proposed that models $f(x)$ nonparametrically using a smoothing spline ANOVA (SS ANOVA) model and $f(y|x)$ parametrically using a conditional Gaussian graphical model (cGGM). Combining the flexibility of the SSANOVA model with the succinctness of the cGGM, this framework allows us to deal with high-dimensional data without assuming a joint Gaussian distribution. A back-fitting estimation procedure is proposed for the cGGM with a computationally efficient approach for the selection of tuning parameters. A geometric inference approach is also developed for edge selection. Asymptotic convergence properties are established for both the parameter and density estimation. The performance of the proposed method is evaluated through extensive simulation studies and real data applications.

E1046: **A comparison of two inconsistency detecting models for network meta-analysis**
Presenter: **Ke Yang**, Beijing University of Technology, China
Co-authors: Lu Qin, Tiejun Tong, Wenlai Guo, Shishun Zhao

The application of network meta-analysis is becoming increasingly widespread, and detecting consistency assumptions has always been one of the most concerned issues. The detection results can serve as a criterion for evaluating the effectiveness of network meta-analysis results. Several methods to detect inconsistency have been proposed. Among them, the design-by-treatment interaction model and the side-splitting models are most commonly used. These two types of models are compared within a frequentist framework. By simple examples of networks with three treatments, it is found that the side-splitting models are specific instances of the design-by-treatment interaction model with additional assumptions. The side-splitting models perform better when these assumptions hold. On the other hand, the design-by-treatment interaction model exhibits robust performance across different data structures. Based on the findings, it is suggested to employ the design-by-treatment interaction model in practical use, with the side-splitting models serving as a supplementary method for inconsistency detection in network meta-analysis.

EO154 Room 201 ADVANCES IN SEMIPARAMETRIC METHODS FOR CAUSAL INFERENCE Chair: Kendrick Li

E0862: **Universal difference-in-differences**
Presenter: **Chan Park**, University of Pennsylvania, United States
Co-authors: Eric Tchetgen Tchetgen

Difference-in-differences (DiD) is a popular method for evaluating real-world policy interventions' causal effects. DiD relies on the parallel trends (PT) assumption to identify the average treatment effect on the treated, which states that the time trends for the average treatment-free potential outcomes are parallel across the treated and control groups. A well-known limitation of the PT assumption is its lack of generalization to causal effects for discrete outcomes and to nonlinear effect measures. Universal Difference-in-Differences (UDiD) based on an alternative assumption to PT is considered for identifying treatment effects for the treated on any scale of potential interest and outcomes of an arbitrary nature. Specifically, the odds ratio equi-confounding (OREC) assumption is introduced, which states that the generalized odds ratios relating the treatment-free potential outcome and treatment are equivalent across time periods. Under the OREC assumption, nonparametric identification for any potential treatment effect on the treated in view is established. Moreover, a consistent, asymptotically linear, and semiparametric efficient estimator is developed for any given treatment effect on the treatment of interest, which leverages recent learning theory. UDiD with simulations and two real-world applications in labour economics and traffic safety evaluation are illustrated.

E1103: **Debiased multivariable Mendelian randomization**
Presenter: **Ting Ye**, University of Pennsylvania, United States

Mendelian randomization (MR) is a method that uses genetic variants as instrumental variables to infer the causal effect of a modifiable exposure on an outcome. Multivariable MR is an extension of standard MR by simultaneously studying multiple exposures. It has two key strengths: it is an effective way to account for horizontal pleiotropy as it can include traits on other pathways as additional exposures, and it can estimate the direct effect of each exposure on the outcome that is not mediated by the other exposures. However, robust multivariable MR faces major statistical and computational challenges. A robust and scalable multivariable MR method, MVMR-dIVW, is proposed which effectively removes the weak instrument bias of the popular multivariable inverse-variance weighted method and can account for balanced horizontal pleiotropy. In conclusion, the results are demonstrated in simulated and real datasets.

E1105: **A novel continuum-of-resistance model and doubly robust for nonresponse adjustment with callback data**
Presenter: **Kendrick Li**, University of Michigan, United States
Co-authors: Xu Shi, Wang Miao

Callback data design is a powerful tool to address missingness not at random in survey sampling. Although the notion of a continuum of resistance on the callback mechanism is widely used in social surveys for leveraging callback data to make nonresponse adjustments, the corresponding statistical analysis literature is relatively sparse. A novel model is proposed that clarifies the underlying assumptions for leveraging the continuum of resistance to make nonresponse adjustments. The proposed model assumes only that the odds ratio functions of the response propensity in the last two captures are equal, conditioning on the other variables. Under this model, the identification was established, and a suite of estimators was developed for the estimation of a general function of the full data law, including a doubly robust one that is consistent if either the callback mechanism or the outcome distribution is specified correctly. The doubly robust estimator with a double machine learning estimator was further extended, which estimates the callback mechanism and the outcome data distribution with flexible machine learning methods. The performance of the proposed estimators with comprehensive simulation studies and an application to ConsumerExpenditure Survey data are demonstrated.

E1127: **Introducing the specificity score: A measure of causality beyond P value**
Presenter: **Wang Miao**, Peking University, China

There has been considerable doubt and debate about using the P value in scientific research in recent years, particularly after its use was banned in several prestigious journals. Much scientific research is concerned with uncovering causal associations. However, the P value, by definition, is a measure of the significance of a statistical association, which could be biased from the causal association of interest and lead to false discoveries due to confounding. A score measuring the specificity of causal associations and a specificity score-based test about the existence of causal effects in the presence of unmeasured confounding will be introduced. Under certain conditions, this approach has controlled type I error and power approaching unity for testing the null hypothesis of no causal effect. This approach is particularly suitable for joint causal discovery with multiple treatments and multiple outcomes, such as gene expression studies, Mendelian randomization and EHR studies. A visualization approach using a specificity map is proposed to communicate all specificity score/test information in a universal and effective manner. Identification and estimation will be briefly covered. Simulations are used for illustration, and an application to a mouse obesity dataset detects potential active effects of genes on clinical traits that are relevant to metabolic syndrome.

EO189 Room 203 RECENT ADVANCES IN NETWORK ANALYSES Chair: Hongmei Zhang

E0743: **Detecting responsible nodes in differential Bayesian networks**
Presenter: **Xianzheng Huang**, University of South Carolina, United States
Co-authors: Hongmei Zhang

To study the roles different nodes play in differentiating Bayesian networks under two states, such as control versus disease, two node-specific scores are formulated to facilitate differential analysis. The first score is motivated by the prediction invariance property of a causal model. The second score results from modifying an existing score constructed for differential analysis of undirected networks. Strategies based on these scores are developed to identify nodes responsible for topological differences between two Bayesian networks. Synthetic data and real-life data from designed experiments are used to demonstrate the effectiveness of the proposed methods in detecting responsible nodes.

E0884: **Causal inference for social network data**
Presenter: **Elizabeth Ogburn**, Johns Hopkins University, United States

Semiparametric estimation and inference are described for causal effects using observational data from a single social network. Our asymptotics allows for the dependence of each observation on a growing number of other units as sample size increases. Both dependencies are allowed due to the transmission of information across network ties and for dependence due to latent similarities among nodes sharing ties. New causal effects are proposed that are specifically of interest in social network settings, such as interventions on network ties and network structure. Our methods are used to reanalyze an influential and controversial study that estimated causal peer effects of obesity using social network data from the Framingham Heart Study; after accounting for network structure, no evidence is found for causal peer effects. Supplementary materials for this article are available online.

E0916: **Comparing dependent directed Gaussian networks**
Presenter: **Hongmei Zhang**, University of Memphis, United States

A Bayesian method is proposed to compare two dependent Gaussian Bayesian (directed) networks with topological ordering unknown. The method unifies topological ordering estimation, network construction, and network comparison. Simulations with different scenarios are used to assess the method. The approach is applied to epigenetic data measured at two-time points to test network differentiation and reconstruct the networks. Theoretical properties, simulation findings, and real applications support the effectiveness of the proposed approach.

E0999: **Unconfoundedness with network interference**
Presenter: **Michael Leung**, UC Santa Cruz, United States

The aim is to study the nonparametric estimation of treatment and spillover effects using observational data from a single large network. A model of network interference is considered that allows for peer influence in outcomes and selection into treatment but requires influence to decay with network distance. In this setting, the network and covariates of all units can be potential sources of confounding, in contrast to existing work that assumes confounding is limited to a known, low-dimensional function of these objects. To estimate the first-stage nuisance functions of the doubly robust estimator, it is proposed to use neural graph networks, which are designed to approximate functions of graph-structured inputs. Under the proposed model of interference, primitive conditions for a network analogue of approximate sparsity are derived, which provides justification for the use of shallow architectures.

EO242 Room 503 BAYESIAN MODELING AND COMPUTATION WITH BEHAVIORAL AND SOCIAL APPLICATIONS Chair: Xiaojing Wang

E1277: **Parallel Markov chain Monte Carlo for Bayesian hierarchical models with big data, in two stages**
Presenter: **Erin Conlon**, University of Massachusetts Amherst, United States
Co-authors: Zheng Wei

Due to the continuing growth of big data sets, new Bayesian Markov chain Monte Carlo (MCMC) parallel computing methods have been created. These methods divide large data sets by observations into subsets. However, many Bayesian hierarchical models have only a small number of parameters that are common to the full data set, with the majority of parameters being group specific. Therefore, techniques that split the full data set by groups rather than by observations are a more natural analysis approach. Such a two-stage Bayesian hierarchical modelling method is adapted and extended. In stage 1, each group is evaluated independently in parallel; the stage 1 posteriors are used as proposal distributions in stage 2, where the full model is estimated. This approach is illustrated using both simulation and real data sets. The results show considerable increases in MCMC efficiency and large reductions in computation times compared to the full data analysis.

E1285: **Variable selection in dynamic item response theory models via Bayes factors with a single MCMC output**
Presenter: **Xiaojing Wang**, University of Connecticut, United States
Co-authors: Jingyu Sun, Liu Yang , Ming-Hui Chen

Item response theory (IRT) models are essential analytic tools used in educational testing. The recent surge in computerized testing enables easier collection of longitudinal data in educational studies, but it brings new challenges for the appropriate analysis of educational testing data. To overcome the limitations of classic IRT models and to handle individually varying and irregularly-spaced longitudinal responses, a dynamic IRT model framework has been employed, and the dynamic changes of ability are further linked with individual characteristics in hierarchical modelling. The algorithm developed only needs one single Markov chain Monte Carlo runs to compute all possible Bayes factors for selecting individual characteristics in the proposed model. Further, the model selection consistency of Bayes factors is verified in both theory and simulations when the Zellner-Siow prior is used. In the end, computerized testing has been applied to illustrate the usage of the proposed model and computational strategies.

E1302: **Variational Bayesian inference for bipartite MMSBM with applications to collaborative filtering**
Presenter: **Panpan Zhang**, Vanderbilt University Medical Center, United States

Motivated by the connections between collaborative filtering and network clustering, a network-based approach is considered to improve rating prediction in recommender systems. A novel Bipartite Mixed-Membership Stochastic Block Model (BM2) with a conjugate prior from the exponential family is proposed. The analytical expression of the model is derived, and a variational Bayesian expectation-maximization algorithm is introduced, which is computationally feasible for approximating the untractable posterior distribution. Extensive simulations are carried out to show that BM2 provides a more accurate inference than standard SBM with the emergence of outliers. Finally, the proposed model is applied to a MovieLens dataset, and it is found that it outperforms other competing methods for collaborative filtering.

E1303: **Managers positive facial expressions and earnings quality**
Presenter: **Wuqing Wu**, Renmin University of China, China
Co-authors: Zhenhan Hong

The relationship between one measure of managers' positive facial expressions and the earnings quality of listed companies is examined. Videos of the annual performance presentations at the Shanghai Roadshow Center are downloaded, and an open-source neural network model is used to analyze the intensity of facial muscle movement associated with smiling. It is found that managers' expression of positivity correlates with less accrued earnings manipulation. This relationship is more pronounced in companies that adjust their accrued earnings upward. Overall, a new measure of the positivity of facial expressions based on neuropsychological theories is constructed, and it is shown that it could be significant nonverbal information for stakeholders to assess firms' performance. Future research could apply this approach in more situations since videos are becoming an increasingly important means of disseminating information.

EO104 Room 506 TRUSTWORTHY MACHINE LEARNING METHODS AND APPLICATIONS Chair: Jie Ding

E0178: **RISE: Robust individualized decision learning with sensitive variables**
Presenter: **Lu Tang**, University of Pittsburgh, United States

RISE, a robust, individualized decision learning framework with sensitive variables, is introduced, where sensitive variables are collectable data essential to the intervention decision. However, their inclusion in decision-making is prohibited due to reasons such as delayed availability or fairness concerns. A naive baseline is to ignore these sensitive variables in learning decision rules, leading to significant uncertainty and bias. To address this, a decision learning framework is proposed to incorporate sensitive variables during offline training but not include them in the input of the learned decision rule during model deployment. Specifically, from a causal perspective, the proposed framework intends to improve the worst-case outcomes of individuals caused by sensitive variables unavailable at the time of decision. Unlike most existing literature that uses mean-optimal objectives, a robust learning framework is proposed by finding a newly defined quantile-or infimum-optimal decision rule. The reliable performance of the proposed method is demonstrated through synthetic experiments and three real-world applications.

E0920: **Integrative regression and factorization of bidimensionally linked matrices**
Presenter: **Eric Lock**, University of Minnesota, United States

Several modern datasets take the form of bidimensionally linked matrices, in which multiple matrices share rows or columns. For example, multiple molecular omics platforms measured for multiple sample cohorts are increasingly common in biomedical studies. A very flexible factorization of such bidimensionally linked data is proposed, allowing for the simultaneous identification of covariate driven-effects and auxiliary structured variation. The approach provides a decomposition of covariate effects and low-rank structure, each of which may be shared across any row sets (e.g., omics platforms) or column sets (e.g., sample cohorts). A structured nuclear norm penalty is used as an objective function, with penalty parameters chosen by random matrix theory. The objective gives the mode of the posterior distribution for an intuitive Bayesian model. The method is applied to pan-omics pan-cancer data from The Cancer Genome Atlas (TCGA), integrating data from several omics platform-seral cancer types.

E0924: **Continuous-time recommender system for implicit feedback**
Presenter: **Xiwei Tang**, University of Virginia, United States
Large volumes of temporal event data are drawing increasing attention in various applications, such as analyzing social media data, healthcare records, online consumption, and product recommendation. Traditional models based on static latent features or discretized time epochs for the recommender system usually fail to capture the essential temporal dynamics in user-item interactions. A novel evolutionary recommender system is proposed by leveraging the temporal mechanism on the continuous-time user-item interactive events. The proposed approach can effectively capture the long- and short-term preferences from the sequential historical data with informative dynamic feature embeddings. An efficient algorithm for learning the model parameters with outstanding scalability and computational effectiveness is developed. Using both synthetic and real-world datasets, the outperformance of the proposed model in learning sequential user behaviours and achieving better predictive power in recommendation is shown.

E0980: **Pruning deep neural networks from a sparsity perspective**
Presenter: **Ganghua Wang**, University of Minnesota, United States
Co-authors: Jie Ding, Yuhong Yang
Recently, deep network pruning has attracted significant attention in order to enable the rapid deployment of AI into small devices with computation and memory constraints. Many deep pruning algorithms have been proposed with impressive empirical success. However, a theoretical understanding of model compression is still limited. One problem is to understand if a network is more compressible than another of the same structure. Another problem is quantifying how much one can prune a network with theoretically guaranteed degradation of accuracy. These two fundamental problems are addressed by using the sparsity-sensitive l_q-norm (0 < q < 1) to characterize compressibility and provide a relationship between the soft sparsity of the network weights and the degree of compression with a controlled accuracy degradation bound. Next, PQ Index (PQI) is proposed to measure the potential compressibility of deep neural networks and use this to develop a Sparsity-informed Adaptive Pruning (SAP) algorithm. Our experiments demonstrate that the proposed adaptive pruning algorithm with a proper choice of hyper-parameters is superior to the iterative pruning algorithms, such as the lottery ticket-based pruning methods, in terms of both compression efficiency and robustness.

EO207 Room 603 CHANGE POINTS DETECTION FOR TIME SERIES Chair: Likai Chen

E0378: **Adaptive two way change points detection**
Presenter: **Likai Chen**, Washington University in Saint Louis, United States
Co-authors: Jiaqi Li
A new detection method is proposed for multiple change points in high-dimensional time series. The method aggregates moving sum (MOSUM) statistics cross-sectionally by an ℓ^2-norm and maximizes them over time. A Two-Way MOSUM statistic is introduced with adaptive window size to account for different break sizes, which can substantially improve the estimation accuracy. The asymptotic theory has been established for the limiting distribution of an ℓ^2-aggregated statistic with varying window sizes for testing the existence of breaks, and the core is to extend a high-dimensional Gaussian approximation theorem to non-stationary and spatial-temporally dependent data generating processes. Consistency results of estimated break numbers, time stamps and sizes of breaks are provided.

E0675: **Adaptive testing in high dimension**
Presenter: **Runmin Wang**, Texas A&M University, United States
Co-authors: Xiaofeng Shao, Yangfan Zhang
A general U-statistic-based approach to adaptive testing for high-dimensional data is introduced. The proposed method extends a recent adaptive test by combining U-statistics for l_q norm of the parameter vector with different $2 \leq q < \infty$, as the larger the q, the better the power against sparse alternatives. For a general parameter vector, it has been proved that the U-statistic for the l_q norm is asymptotically normal under mild regularity conditions. More importantly, such U-statistics for different q are still asymptotically independent, which has already been shown for the specific problems discussed previously. Further, a new test is developed only using subsamples with monotone indices to reduce the computational cost with mild efficiency loss. It was proved that the new method could speed up the calculation by a lot with mild efficiency loss. An application of the proposed test to change point detection will also be discussed. Simulation studies indicate that the new method is powerful against both dense and sparse alternatives for numerous problems.

E0939: **Change point analysis with irregular signals: When did the COVID-19 pandemic start?**
Presenter: **Wei Biao Wu**, University of Chicago, United States
The focus is on testing and estimating change points where signals after the change point can be highly irregular. The setting substantially differs from the existing literature, assuming signals are piecewise constant or vary smoothly. A two-step approach is proposed to estimate the location of the change point effectively in. The first step consists of a preliminary estimation of the change point that allows for obtaining unknown parameters in the second step. In the second step, a new procedure is used to determine the position of the change point. It is shown that the optimal OP(1) rate of convergence of the estimated change point to the true change point can be obtained. The method is applied to analyze health time series data and estimate 7 December 2019 as the starting date of the COVID-19 pandemic.

E0165: **Change points detection in VAR models**
Presenter: **Siddhartha Chib**, Washington University in Saint Louis, United States
A Bayesian method for conducting inference on possible single or multiple change points in VAR models is presented. Under a conjugate prior to the parameters of the VAR model, it is shown that under regularity conditions, the posterior distribution of the change-point location (in a model with a single change-point) concentrates on the true change point as the sample sizes become large. By a novel method of proof, this result is extended to a family of non-conjugate priors on the VAR parameters. VAR models are considered with multiple change points where results on posterior consistency of change-point determination are conditioned on knowledge of other change points. Establishing joint posterior consistency of multiple change points remains an open problem. The change-point detection methodology is applied to several problems and shows that change points are common in typical VAR applications.

EO009 Room 604 RECENT ADVANCES IN FINANCIAL ECONOMETRICS Chair: Toshiaki Watanabe

E0234: **Tail risk forecasting of realized volatility CAViaR models**
Presenter: **Cathy W-S Chen**, Feng Chia University, Taiwan
Co-authors: Hsiao-Yun Hsu, Toshiaki Watanabe

A new class of RES-CAViaR (conditional autoregressive value-at-risk) models that incorporate daily realized volatility and expected shortfall (ES) is proposed to forecast VaR and ES simultaneously. Weekly and monthly realized volatilities in the proposed model are further considered to approximate a long-memory process. The Bayesian adaptive Markov chain Monte Carlo approach is employed to estimate all unknown parameters and to jointly predict daily VaR and ES over a 4-year out-of-sample period, including the COVID-19 pandemic. The results show that the realized CAViaR-type models outperform in terms of three backtests, four loss-function criteria, and ES measurement at the 1% level.

E0571: **Time-varying parameter local projections with stochastic volatility**
Presenter: **Jouchi Nakajima**, Hitotsubashi University, Japan

The local projection method has been widely used as a promising framework for computing impulse responses. In the previous literature, a time-varying version of the local projection has been proposed, but it does not address the time-varying variance of errors. Ignoring a possible time variation in the error variance could cause a severe bias in the time-varying impulse responses. To overcome it, the aim is to propose the time-varying parameter local projections with stochastic volatility. A Bayesian efficient estimation method is developed to analyze the proposed model. The application to returns of financial variables is provided.

E0647: **Analyzing intraday variation in price impact: A Bayesian SVAR approach with stochastic volatility estimation**
Presenter: **Makoto Takahashi**, Hosei University, Japan

The aim is to analyze the intraday variation in the short- and long-term price impact of market orders, limit orders, and cancellations using a structural vector autoregression (SVAR) model. While Bayesian estimation using sign restrictions has been effective in parameter estimation in SVAR models, there are issues with parameter uniqueness. Alternative methods like maximum likelihood estimation and generalized method of moments have been proposed to address this. A new Bayesian estimation method is applied that considers the stochastic volatility of errors to estimate the model parameters. The method allows the unique identification of the parameters without being affected by order of the variables by imposing sign conditions on the variables in addition to the heteroskedasticity of the variables. The advantage is that the sign conditions can be easily verified from the posterior distribution of the estimated parameters. This estimation method has not been applied to high-frequency order book data, but the use of a large number of observations allows the model to be estimated every few minutes to tens of minutes and examine the intraday variation. The model simultaneously analyzes the variation in price impact and volatility by modelling and estimating the stochastic variation of both price changes and orders.

E0760: **A realized multi-factor regression model with realized stochastic volatility**
Presenter: **Tsunehiro Ishihara**, Takasaki City University of Economics, Japan

Estimation of high-dimensional stochastic volatility models tends to be computationally expensive. A stochastic volatility model is proposed that can be computed in parallel, even in high dimensions, and does not increase the computational cost. Realized covariance is computed from indices' high-frequency data of market, size, and value quasi-factors. Using them, a time-varying coefficient regression model or a low-dimensional stochastic volatility model is estimated to forecast high-dimensional volatility. 33-dimensional Japanese sector indices data are applied.

EO239 Room 605 ANALYSIS CHALLENGES FOR COMPLEX FEATURED DATA Chair: Wenqing He

E0434: **Semiparametric additive time-varying coefficients model for longitudinal data with censored time origin**
Presenter: **Yanqing Sun**, University of North Carolina at Charlotte, United States
Co-authors: Qiong Shou , Peter Gilbert, Fei Heng, Xiyuan Qian

Statistical analysis of longitudinal data often involves modelling treatment effects on clinically relevant longitudinal biomarkers since an initial event (the time origin). In some studies, including preventive HIV vaccine efficacy trials, some participants have biomarkers measured starting at the time origin. In contrast, others have biomarkers measured starting later with the time origin unknown The semiparametric additive time-varying coefficient model is investigated where the effects of some covariates vary nonparametrically with time while the effects of others remain constant Weighted profile least squares estimators coupled with kernel smoothing are developed The method uses the expectation maximization approach to deal with the censored time origin The Kaplan-Meier estimator and other failure time regression models, such as the Cox model, can be utilized to estimate the distribution and the conditional distribution of left-censored event time related to the censored time origin Asymptotic properties of the parametric and nonparametric estimators and consistent asymptotic variance estimators are derived A two-stage estimation procedure for choosing weight is proposed to improve estimation efficiency Numerical simulations are conducted to examine the finite sample properties of the proposed estimators The method is applied to analyze data from the Merck 023/HVTN 502 Step HIV vaccine study.

E1178: **Deep generative estimation of conditional survival function**
Presenter: **Xingqiu Zhao**, The Hong Kong Polytechnic University, Hong Kong
Co-authors: Xingyu Zhou, Wen Su, Changyu Liu, Yuling Jiao, Jian Huang

A deep generative approach is proposed to the nonparametric estimation of conditional survival and hazard functions with censored data. The key idea of the proposed method is first to learn a conditional generator for the joint conditional distribution of the observed time and censoring indicator given covariates and then construct the Kaplan-Meier and Nelson-Aalen estimators based on this conditional generator for conditional hazard and survival functions. The method combines ideas from the recently developed deep generative learning and classical nonparametric estimation in survival analysis. The convergence properties of the generative nonparametric estimators are established. The numerical studies validate the proposed method.

E1193: **Semiparametric methods for left-truncated and right-censored survival data with covariate measurement error**
Presenter: **Grace Yi**, University of Western Ontario, Canada

Biased samples caused by left-truncation (or length-biased sampling) and measurement error often accompany survival analysis. While such data frequently arise in practice, little work has been available to address these features simultaneously. Valid inference methods are explored for handling left-truncated and right-censored survival data with measurement error under the widely used Cox model. First, a flexible estimator is exploited for the survival model parameters, which does not require the specification of the baseline hazard function. An augmented nonparametric maximum likelihood estimator is further developed to improve efficiency. Asymptotic results are established, and the efficiency and robustness issues for the proposed estimators are examined. The proposed methods enjoy appealing features in that the distributions of the covariates and the truncation times are left unspecified. Numerical studies are reported to assess the finite sample performance of the proposed methods.

E1194: **Identification of survival relevant genes with measurement error in gene expression incorporated**
Presenter: **Wenqing He**, University of Western Ontario, Canada

Modern gene expression technologies, such as microarray and the next generation RNA sequencing, enable simultaneous measurement of expressions of a large number of genes and therefore represent important tools in personalized medicine research for improving patient survival prediction accuracy. However, survival analysis with gene expression data can be challenging due to the high dimensionality. Proper identification of survival-relevant genes is thus imperative for building suitable prediction models. In spite of the fact that gene expressions are typically subject to measurement errors introduced from the complex experimental procedure, the issue of measurement error is often ignored in survival gene identifications. The effect of measurement error on the identification of survival-relevant genes is explored under the accelerated failure time model setting. Survival-relevant genes are identified by regularizing the weighted least square estimator with the adaptive LASSO penalty. The simulation-extrapolation method is incorporated to adjust for the impact of measurement error on gene identification. The performance of the proposed method is assessed by simulation studies, and the utility of the proposed method is illustrated by a real data set collected from the diffuse large-B-cell lymphoma study. The results show that the proposed method yields better prediction models than traditional methods that ignore gene expression measurement errors.

EO123 Room 606 NONSTATIONARY AND HIGH-DIMENSIONAL TIME SERIES: SOLUTIONS FOR CHALLENGES Chair: Ansgar Steland

E0816: **Estimation of the strongly spiked eigenstructure in high-dimensional settings**
Presenter: **Kazuyoshi Yata**, University of Tsukuba, Japan
Co-authors: Aki Ishii, Makoto Aoshima

High-dimensional data often have a low-rank structure which contains strongly spiked eigenvalues. The problem of estimating the strongly spiked eigenstructure in high-dimensional situations is considered. First, the conventional PCA is considered to estimate the structure, showing that the estimation holds consistency properties under severe conditions. The conventional PCA is heavily subjected to noise. Recently, consistent estimators of the strongly spiked eigenvalues and eigenvectors have been given by developing a new PCA method called the Automatic Sparse PCA (A-SPCA) methodology. To remove the noise, the A-SPCA is applied and propose a new estimation of the strongly spiked eigenstructure. The proposed estimation by the A-SPCA holds the consistency properties under mild conditions and effectively improves the conventional PCA's error rate effectively. Finally, the performance of the proposed estimation is investigated in actual data analyses.

E0952: **Inference and change-point testing of high-dimensional spectral density matrices: Beyond spectral averages**
Presenter: **Ansgar Steland**, RWTH Aachen University, Germany

A flexible approach is studied to analyze high-dimensional nonlinear time series based on linear statistics calculated from spectral average statistics of bilinear forms and nonlinear transformations of lag-window (i.e. band-regularized) spectral density matrix estimators. That class of statistics includes, among others, smoothed periodograms, nonlinear statistics such as coherency, long-run-variance estimators and statistics related to factorial effects as special cases. The novel class of spectral averages of nonlinear functions of the spectral density matrix is introduced. Big data settings are considered by studying sparse sampling. Optimal rate Gaussian approximations are derived for non-stationary nonlinear time series. For change-testing (self-standardized), CUSUM statistics are examined. Further, a specific wild bootstrap procedure is proposed to estimate critical values. A simulation study studies finite sample properties of the proposed self-standardized CUSUM bootstrap testy. The results indicate that the procedure keeps the nominal significance level and performs well under a change-point alternative. The approach is illustrated by analyzing SP500 financial returns from 2016 to 2022, considering a certain well-diversified buy-and-hold portfolio investing 50% in the minimal-variance portfolio, 25% in health care, 10% in Tech firms and 15% in industrial companies. The results show that all significant change points correspond to real market turmoils.

E0987: **Unrestricted maximum likelihood estimation of multivariate realized volatility models**
Presenter: **Jan Vogler**, Ruhr Universitaet Bochum, Germany
Co-authors: Vasyl Golosnoy

The popular conditional autoregressive Wishart (CAW) model for dynamics of realized covariance matrices provides a flexible parametrization. However, the number of parameters grows quadratically with the number of assets, which causes enormous computational difficulties in higher dimensions. Therefore, its unrestricted maximum likelihood (ML) estimation up to now has been conducted only for small portfolios with around five assets. It is elaborated on unrestricted ML estimation of the CAW model in higher dimensions abound 30 assets, which is a sufficient number for portfolio diversification. This can be done by providing various explicit analytical results for computing the gradient for log-likelihood optimization.

E0589: **Rank tests for randomness against time-varying MA alternative**
Presenter: **Junichi Hirukawa**, Niigata University, Japan

The idea of the problem of testing randomness against ARMA alternatives to a class of locally stationary processes introduced by Dahlhaus is extended. The linear serial rank statistics are used, and the notion of the contiguity by LeCam for the testing problem is applied. Under the null hypothesis, the joint asymptotic normality of the proposed rank test statistics and the log-likelihood ratio is established using the local asymptotic normal property. Then, applying LeCam's third lemma, the asymptotic normality of the test statistic under the alternative is automatically derived.

EO172 Room 701 RECENT DEVELOPMENTS IN DEGRADATION ANALYSIS AND RELATED TOPICS II Chair: Chien-Yu Peng

E0930: **Planning of an accelerated degradation test**
Presenter: **I-Chen Lee**, National Cheng Kung University, Taiwan

Accelerated degradation tests (ADTs) are widely used to access the lifetime information of highly reliable products. To obtain a more accurate prediction of lifetime information, how to design an efficient experiment under a limited budget is a critical issue for reliability analysts. Much literature addressed this problem and indicated that a two-level design is an optimum strategy for an ADT plan. Considering easier operating conditions for experimenters, most literature developed the optimum designs under the assumption that the numbers of measurements and the duration between two inspections within a degradation path are equal for all testing stress levels. However, some real applications were conducted under the operating conditions that the numbers of measurements and sampling frequencies were different for all stress levels. Based on the exponential dispersion (ED) degradation model, the optimum planning under various constraints of operating conditions is determined so that the asymptotic variance of a prediction can be minimized.

E0959: **A new Phase II change-point detection control chart for monitoring and diagnostics of linear profiles**
Presenter: **Longcheen Huwang**, National Tsing Hua University, Taiwan

A new Phase II control chart, which is based on the change-point model and combined with the exponentially weighted moving average (EWMA) mechanism, is proposed to monitor general linear profiles. The new control chart can be used to monitor general linear profiles when the true in-control parameters are unknown and only a few historical data are available. In addition, when the chart triggers an out-of-control signal, not only can it estimate the location of the change point, it can also identify which of the parameters have changed as well as the change directions. Using Monte-Carlo simulations, the proposed chart is shown to be effective and has good diagnostic performance. Furthermore, the simulation results show that the proposed chart has better performance than the existing charts in most of the out-of-control scenarios considered. An example is used to illustrate how the proposed chart can be implemented in practical applications.

E0995: **Analysis of zero-increment degradation data**
Presenter: **Yi-Fu Wang**, National Cheng Kung University, Taiwan

For high-reliability products, degradation analysis is a commonly used tool to predict the product's lifetime. The gamma process model is the most widely utilized degradation model to fit the monotonic degradation path. However, as the monotonic degradation path appears in some zero-increment cases, it causes the gamma process model cannot be adopted to implement the degradation analysis. Thus, some nondecreasing gamma process models that incorporate the mixture probability distribution to deal with the zero-increment degradation data are proposed. The homogeneous and nonhomogeneous properties for the probability of occurrence of zero-increment are considered. In addition, it is also considered that the smallest limitation of measuring instruments in recording observations exists. Simulation studies and real examples are carried out to illustrate the proposed models.

E1317: **The first-passage-time moments for Hougaard Process and its Birnbaum-Saunders approximation**
Presenter: **Yi-Shain Dong**, National Central University, Taiwan
Co-authors: Tsai-Hung Fan, Chien-Yu Peng

Hougaard processes, which include gamma and inverse Gaussian processes as special cases, as well as the moments of the corresponding first-passage-time (FPT) distributions, are commonly used in many applications. Because the density function of a Hougaard process involves an intractable infinite series, the Birnbaum-Saunders (BS) distribution is often used to approximate its FPT distribution. The aim is to derive the finite moments of FPT distributions based on Hougaard processes and provide a theoretical justification for BS approximation regarding convergence rates. Further, it is shown that the first moment of the FPT distribution for a Hougaard process approximated by the BS distribution is larger and provides a sharp upper bound for the difference using an exponential integral. The conditions for convergence coincidentally elucidate the classical convergence results of Hougaard distributions. Some numerical examples are proposed to support the validity and precision of the theoretical results.

EO205 **Room 702** **TOPICS IN MICROBIOME DATA ANALYSIS** **Chair: Julia Fukuyama**

E0633: **Statistical methods to analyze phylogenetic trees with non-identical leaf sets**
Presenter: **Maria Valdez Cabrera**, University of Washington, United States

Phylogenetic trees are frequently used to describe the evolutionary history of a set of microorganisms. Different genes shared by these may differ in their evolutionary histories, motivating methods to analyze collections of trees. To allow comparisons between phylogenetic trees, a non-Euclidean metric space (BHV space) that accounts for the discrete branching structure of each tree and allows the branch lengths to vary continuously was introduced in 2001. Unfortunately, only trees with identical leaf sets are elements in this space. In practice, this might not be reasonable. Some microorganisms may not carry a particular gene, or a gene may not be detected in a sample due to laboratory and technical artefacts. Motivated by the path continuity of BHV geodesics, a metric space is proposed for phylogenetic trees with potentially non-identical leaf sets. An algorithm is introduced to compute the distance in our metric space and discuss the use of the Frechet mean as a potential summary for a tree collection. The long-term goal is to create statistical tools to analyze a collection of phylogenetic trees with non-identical leaf sets. This is joint work with Amy Willis.

E0869: **Generalized matrix decomposition regression and inference for two-way structured data**
Presenter: **Timothy Randolph**, Fred Hutchinson Cancer Research Center, United States
Co-authors: Yue Wang, Ali Shojaie, Parker Knight, Jing Ma

Two-way structured data arise naturally in many applications, including microbiome studies. Characteristics of these data are (i) structured relationships among the variables, which may be informed by extrinsic information, and/or (ii) non-independent and non-Euclidean relationships among observations. For modelling data of this type, a penalized regression framework is proposed that exploits the Generalized Matrix Decomposition (GMD), a natural extension of classical dimension reduction methods such as principal component analysis (PCA) and generalized PCA. The GMD of a matrix accounts for the prescribed structure among its columns (variables) and rows (observations). The GMD regression approach includes efficient estimation, valid inference, and exploratory graphics in a GMD biplot.

E0913: **Distance-based run tests from complex high-dimensional data**
Presenter: **Debashis Mondal**, Washington University in St Louis, United States
Co-authors: Arpita Mukherjee

Distance-based two-sample run tests and computation for analyzing complex, high-dimensional data that arise from compositions, trees, graphs, or networks, are discussed. The distances considered are all non-Euclidean. They could be either non-metric dissimilarities that do not satisfy any triangular inequalities or even just discrete numbers, but they all arise from conditionally positive definite kernels. Examples of distances include the Bray- Curtis dissimilarity, the Unifrac metric, the Aitchison distance, graph kernels, spectral distances, and other distances based on optimal transport problems. The run test is constructed by counting runs along the shortest Hamiltonian paths/ loops of the data points. These run tests are shown to be exact, distribution-free, and consistent as the dimension of the data points goes to infinity, but the total number of data points is fixed. Asymptotic results are provided when the number of data points goes to infinity by expanding previous work. The method is illustrated through a simulation study with the Ewens sampling formula and a suite of dissimilarity measures. Two applications are further presented; one concern checking homogeneity across the forest dynamic plot of Barro Colorado Island, Panama, and the other analyzes 16S microbial community data. The work is supported by an NSF grant from the Division of Mathematical Sciences.

E0857: **Logistic-normal multinomial mediation analysis of microbiome community profiles**
Presenter: **Kris Sankaran**, University of Wisconsin, United States

To design microbiome-based therapies, it is necessary to understand the consequences of interventions on the microbiome. Though randomized controlled trials have demonstrated that various interventions can impact microbiome composition, it is challenging to distinguish between potential mechanisms without further study of mediating factors. Building from work on mediation analysis in the compositional setting, a framework for Logistic-Normal Multinomial mediation analysis is introduced where the response of interest is a microbiome profile. Instantiations of this framework that posit specific zero-inflation, latent factor, and longitudinal structure are described. A zero-inflated quantiles-based simulation procedure is developed to guide model selection and calibrate inferences. It is illustrated how our workflow can discriminate between candidate causal mechanisms in a study of the effects of a mindfulness-based intervention on the microbiome. Our R package, LNMmediation, can be found online.

EO070 Room 703 STATISTICAL ANALYSIS FOR DATA WITH COMPLEX STRUCTURES Chair: Binyan Jiang

E0298: **Directed community detection with network embedding**
Presenter: **Jingnan Zhang**, University of Science and Technology of China, China

Community detection in network data aims at grouping similar nodes sharing certain characteristics together. Most existing methods focus on detecting communities in undirected networks, where the similarity between nodes is measured by their node features and whether they are connected. A novel method is proposed to conduct network embedding and community detection simultaneously in a directed network. The network embedding model introduces two sets of vectors to represent the out- and in-nodes separately and thus allows the same nodes to belong to different out- and in-communities. The community detection formulation equips the negative log-likelihood with a novel regularization term to encourage community structure among the nodes representations and thus achieves better performance by jointly estimating the nodes embeddings and their community structures. To tackle the resultant optimization task, an efficient alternative updating scheme is developed. More importantly, the asymptotic properties of the proposed method are established in terms of both network embedding and community detection, which are also supported by numerical experiments on some simulated and real examples.

E0386: **Factor modelling for clustering high-dimensional time series**
Presenter: **Bo Zhang**, University of Science and Technology of China, China

A new unsupervised learning method is proposed for clustering a large number of time series based on a latent factor structure. Each cluster is characterized by its own cluster-specific factors in addition to some common factors which impact all the time series concerned. The setting also offers the flexibility that some time series may not belong to any clusters. The consistency with explicit convergence rates is established for the estimation of the common factors, the cluster-specific factors, and the latent clusters. A numerical illustration with both simulated data and a real data example is also reported. As a spin-off, the proposed new approach also significantly advances the statistical inference for a previous factor model.

E0432: **Longitudinal elastic shape analysis of surfaces**
Presenter: **Yuexuan Wu**, University of Washington, United States

Over the past 30 years, magnetic resonance imaging (MRI) has become a ubiquitous tool for accurately visualizing the development of subcortical structures in the brain. However, the quantification of complex subcortical structures is still in its infancy due to challenges in shape extraction, representation, and modelling. A simple and efficient framework of longitudinal elastic shape analysis (LESA) is introduced for subcortical structure surfaces. Integrating ideas from elastic shape analysis of static surfaces and statistical modelling of sparse longitudinal data, LESA provides a set of tools for systematically quantifying longitudinal changes in subcortical surface shapes from raw structural MRI data. LESA can efficiently represent complex subcortical structures using a small number of basis functions and can accurately predict the spatiotemporal shape changes of the surfaces. Besides, by applying LESA to analyze three longitudinal neuroimaging data sets, its wide applications are showcased in estimating continuous shape trajectories, building life-span growth patterns, and comparing shape differences among different groups.

E1144: **A two-way heterogeneity model for dynamic networks**
Presenter: **Binyan Jiang**, The Hong Kong Polytechnic University, Hong Kong

Analysis of networks that evolve dynamically requires the joint modelling of individual snapshots and time dynamics. A new flexible two-way heterogeneity model towards this goal is proposed. The new model equips each network node with two heterogeneity parameters, one to characterize the propensity to form ties with other nodes statically and the other to differentiate the tendency to retain existing ties over time. With n observed networks each having p nodes, a new asymptotic theory is developed for the maximum likelihood estimation of $2p$ parameters when $np \to \infty$. The negative log-likelihood function's global non-convexity is overcome by its local convexity, and propose a novel method of moment estimator as the initial value for a simple algorithm that leads to the consistent local maximum likelihood estimator (MLE). To establish the upper bounds for the estimation error of the MLE, a new uniform deviation bound is derived, which is of independent interest. The theory of the model and its usefulness are further supported by extensive simulation and a data analysis examining the social interactions of ants.

EO093 Room 704 ADVANCED DEVELOPMENTS IN COMPLEX DATA ANALYSIS AND EXPERIMENTAL DESIGN Chair: Ming-Chung Chang

E0221: **Inference for the dimension of a regression relationship using pseudo-covariates**
Presenter: **Shih-Hao Huang**, National Central University, Taiwan
Co-authors: Kerby Shedden, Hsin-wen Chang

The main goal of data analysis using dimension reduction methods is to summarize how the response is related to the covariates through a few linear combinations. One key issue is determining the minimal number of relevant covariate combinations, which is the dimension of the sufficient dimension reduction (SDR) subspace. The purpose is to propose an easily-applied approach to conduct inference for the dimension of the SDR subspace based on augmentation of the covariate set with simulated pseudo-covariates. Applying the partitioning principle to the possible dimensions, rigorous sequential testing is used to select the dimensionality by comparing the strength of the signal arising from the actual covariates to that appearing to arise from the pseudo-covariates. It is shown that under a "uniform direction" condition, this approach can be used in conjunction with several popular SDR methods, including sliced inverse regression. In these settings, the test statistic asymptotically follows a beta distribution and therefore is easily calibrated. Moreover, the sequential testing's family-wise type I error rate is rigorously controlled. Simulation studies and an analysis of newborn anthropometric data demonstrate the robustness of the proposed approach and indicate that the power is comparable to or greater than the alternatives.

E0222: **Generating optimal order-of-addition two-level factorial designs**
Presenter: **Shin-Fu Tsai**, National Taiwan University, Taiwan

In many industrial and scientific studies, efficient experimental designs can extract as much information as possible from the observed data under resource constraints. This study will introduce a new class of optimal designs for order-of-addition two-level factorial experiments in which both component addition orders and component levels can be varied over treatments. The proposed designs can be generated by combining order-of-addition orthogonal arrays and two-level orthogonal arrays. Therefore, pairwise order effects and component main effects can be estimated with optimal efficiency. Several construction methods will be introduced to generate these kinds of optimal designs. A drug combination study will also be presented to show that the proposed designs can be practical for real-world studies.

E0645: **Functional-input Gaussian processes with applications to inverse scattering problems**
Presenter: **Chih-Li Sung**, Michigan State University, United States

Surrogate modelling based on Gaussian processes (GPs) has received increasing attention in the analysis of complex problems in science and engineering. Despite extensive studies on GP modelling, developments for functional inputs are scarce. Motivated by an inverse scattering problem in which functional inputs representing the support and material properties of the scatterer are involved in the partial differential equations, a new class of kernel functions for functional inputs is introduced for GPs. Based on the proposed GP models, the asymptotic convergence properties of the resulting mean squared prediction errors are derived, and the finite sample performance is demonstrated by numerical examples. In the application of inverse scattering, a surrogate model is constructed with functional inputs, which is crucial to recover the reflective index of an inhomogeneous isotropic scattering region of interest for a given far-field pattern.

E0352: **Sufficient dimension reduction for Poisson regression**
Presenter: **Jianxuan Liu**, Syracuse University, United States

Poisson regression is popular and commonly employed to analyze the frequency of occurrences in a fixed amount of time. In practice, data collected from many scientific disciplines are often high-dimensional. Sufficient dimension reduction (SDR) is known to be an effective cure for its advantage of making use of all available covariates. Existing SDR techniques for a continuous or binary response do not naturally extend to count response data. Detecting the dependency between the response variable and the covariates is challenging due to the curse of dimensionality. To bridge the gap between SDR and its applications in count response models, an efficient estimating procedure is developed to recover the central subspace by estimating a finite-dimensional parameter in a semiparametric model. The proposed model is flexible and does not require model assumption on the conditional mean or multivariate normality assumption on covariates. The resulting estimators achieve optimal semiparametric efficiency without imposing linearity or constant variance assumptions. The finite sample performance of the estimators is examined via simulations, and the proposed method is further demonstrated in the baseball hitter example and pathways to desistance study.

EO079 Room 705 STATISTICAL CHALLENGES FOR COMPLEX BRAIN SIGNALS AND IMAGES Chair: Michele Guindani

E1171: **A predictor-informed multi-subject Bayesian approach for dynamic functional connectivity**
Presenter: **Michele Guindani**, University of California Los Angeles, United States

Time-Varying Functional Connectivity (TVFC) investigates dynamic interactions between brain regions over fMRI experiments. These changes can be modulated by underlying physiological mechanisms such as attention or cognitive effort. A multi-subject Bayesian framework is proposed to estimate dynamic functional networks based on time-varying exogenous physiological covariates. A non-homogeneous hidden Markov model is used to classify fMRI time series into latent neurological states, allowing for the estimation of recurrent connectivity patterns and the sharing of networks among subjects. The model also assumes sparsity in network structures via shrinkage priors, with edge selection achieved through a multi-comparison procedure for shrinkage-based inferences with Bayesian false discovery rate control. The framework is applied to a resting-state experiment with concurrent pupillometry measurements, revealing the effects of changes in pupil dilation on the subjects' propensity to change connectivity states.

E1186: **Accurate individualized functional brain connectivity and topography via ICA with empirical population priors**
Presenter: **Amanda Mejia**, Indiana University, United States

Independent component analysis (ICA) is often applied to functional MRI data to estimate functional topography and connectivity (FC). However, subject-level ICA results are typically too noisy due to high noise levels to be practically useful. Hierarchical Bayesian ICA models leverage information shared across subjects to improve estimation efficiency, but they have several limitations. Functional connectivity template ICA, a single-subject Bayesian ICA model, is proposed using empirical population priors on spatial topology and functional connectivity between spatial components. These priors can be derived from large fMRI databases or holdout data. Compared with hierarchical models, the proposed approach is computationally convenient, allows for more complex model formulations, and is potentially clinically applicable. This approach is validated through simulation studies and data from the Human Connectome Project and the Midnight Scan Club.

E1191: **A Bayesian semi-parametric model for 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 monitor brain activity. 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 courser-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. A fully Bayesian semi-parametric model is proposed to analyze fNIRS data. Two defining features delineating my model from standard methods are using a time-varying autoregression component to handle the temporal correlation and B-splines bases to model low-frequency drift and motion artefacts. Simulation studies show that this Bayesian model easily handles motion artefacts and results in better statistical properties than the most widely used model, referred to as the AR-IRLS model. Then the model is fitted to two real datasets.

E0482: **Large-scale optimal transport and its application in biomedical research**
Presenter: **Jingyi Zhang**, Tsinghua University, China

Optimal transport has been one of the most exciting subjects in mathematics, starting from the 18th century. As a powerful tool to transport between two probability measures, optimal transport methods have been nowadays in a remarkable proliferation of modern data science applications. Various computational tools have been developed in the recent decade to meet the big data challenges to accelerate the computation for optimal transport methods. A projection-based optimal transport method is presented. Then its real-world applications in biomedical research are discussed.

EO197 Room 708 STATISTICAL METHODOLOGIES FOR INFINITE DIMENSIONAL DATA Chair: Subhra Sankar Dhar

E0641: **Model averaging for global Frechet regression**
Presenter: **Daisuke Kurisu**, The University of Tokyo, Japan
Co-authors: Taisuke Otsu

The analysis of non-Euclidean complex data is gaining popularity in various domains of data science. In a seminal paper, the concept of regression analysis was generalized to accommodate non-Euclidean response objects. On the other hand, model averaging has a long-standing history in conventional regression analysis and is extensively utilized in the statistical literature. The notion of model averaging to global Frechet regressions is extended, and the optimal property of cross-validation in selecting the averaging weights for minimizing the final prediction error is established. A simulation study demonstrates the excellent out-of-sample predictions achieved by the proposed method.

E0811: **Finite-dimensional realizations for stochastic PDEs**
Presenter: **Suprio Bhar**, Indian Institute of Technology Kanpur, India

Stochastic PDEs have become an important modelling tool in describing the evolution of complex systems. These equations are formulated in infinite-dimensional spaces, and the solutions are usually measure/distribution-valued stochastic processes. Two special classes of Stochastic PDEs, which can be reformulated into finite-dimensional Stochastic Differential Equations, are considered, and the existence and uniqueness of solutions are studied. Examples of these Stochastic PDEs appear in certain interest rate models and the evolution of a system of particles.

E0863: **Statistical inference for mean function of longitudinal imaging data over complicated domains**
Presenter: **Jie Li**, Renmin University of China, China

Motivated by longitudinal imaging data, which possesses inherent spatial and temporal correlation, a novel procedure is proposed to estimate its mean function. The functional moving average is applied to depict the dependence among temporally ordered images. Flexible bivariate splines over triangulations are used to handle the irregular domain of images, which is common in imaging studies. Both global and local asymptotic properties of the bivariate spline estimator for mean function are established with simultaneous confidence corridors (SCCs) as a theoretical byproduct. Under some mild conditions, the proposed estimator and its accompanying SCCs are shown to be consistent and oracle efficient as if all images were entirely observed without errors. The finite sample performance of the proposed method through Monte Carlo simulation experiments strongly corroborates the asymptotic theory. The proposed method is illustrated by analyzing two seawater potential temperature data sets.

E0825: **Concurrent object regression**
Presenter: **Satarupa Bhattacharjee**, Pennsylvania State University, United States
Co-authors: Hans-Georg Mueller

Modern-day problems in statistics often face the challenge of exploring and analyzing complex non-Euclidean object data that do not conform to vector space structures or operations. Examples of such data objects include covariance matrices, graph Laplacians of networks, and univariate probability distribution functions. A new concurrent Frechet regression model is proposed to characterize the time-varying relation between an object in a general metric space (as a response) and a multivariate real-valued vector (as a predictor). Concurrent regression has been a well-studied area of research for Euclidean predictors and responses, with many important applications for longitudinal studies and functional data. Generalized versions of global least squares regression and locally weighted least squares smoothing, both in the context of concurrent regression for responses situated in general metric spaces, are developed. Estimators that can accommodate sparse and/or irregular designs are proposed. Consistency results are demonstrated for the sample estimates towards appropriate population targets along with the corresponding rates of convergence. The proposed models are illustrated with mortality data and resting state functional Magnetic Resonance Imaging data (fMRI) as responses.

EO181 Room 709 STATISTICAL LEARNING ON COMPLEX DATA Chair: Ting Li

E0223: **A general pairwise comparison model for extremely sparse networks**
Presenter: **Ruijian Han**, The Hong Kong Polytechnic University, China

Statistical estimation using pairwise comparison data is an effective approach to analyzing large-scale sparse networks. A general framework is proposed to model the mutual interactions in a network which enjoys ample flexibility in terms of model parameterization. Under this setup, the study shows that the maximum likelihood estimator for the latent score vector of the subjects is uniformly consistent under a near-minimal condition on network sparsity. This condition is sharp in terms of the leading order asymptotics describing sparsity. This analysis uses a novel chaining technique and illustrates an important connection between graph topology and model consistency. The results guarantee that the maximum likelihood estimator is justified for estimation in large-scale pairwise comparison networks where data are asymptotically deficient. Simulation studies are provided in support of the theoretical findings.

E0248: **Wasserstein generative regression**
Presenter: **Jian Huang**, The Hong Kong Polytechnic University, China

A Wasserstein generative regression (WGR) approach is proposed to learn a general regression function nonparametrically using deep neural networks. WGR is based on an objective function constructed by regularizing the usual least squares loss with the Wasserstein distance between the distribution of the regression function and the data distribution. WGR learns a general regression function that can also serve as a generator for sampling from the conditional distribution of the response given the predictor. This is different from the usual regression methods that only learn the conditional mean or conditional quantile functions of the response given the predictor. Another attractive feature of WGR is that it can easily handle high-dimensional responses and predictors. Some preliminary results on the consistency and non-asymptotic error bounds of WGR under appropriate conditions are presented. Extensive numerical experiments are also conducted to demonstrate the advantages of WGR.

E0370: **Two-way node popularity model for directed and bipartite networks**
Presenter: **Ting Li**, Hong Kong Polytechnic University, Hong Kong

Community detection for directed and bipartite networks has raised lots of interest in recent decades. While many of the existing results introduced block-wise structure, most of them have the restriction that nodes in the same community behave identically or change uniformly in all communities. However, the heterogeneous node popularity is widely observed both in undirected and directed networks and has been studied under undirected scenarios. Motivated by the variability of node popularity in empirical directed networks, a novel probabilistic framework is proposed for directed network community detection, called the two-way node popularity model (TNPM). To fit the proposed model, the Rank One Approximation Algorithm (ROA) and establish the consistency of ROA is developed for community detection. In addition, an alternative computationally efficient algorithm, called Two-Stage Divided Cosine Algorithm (TSDC), is proposed to fit large-scale networks. Extensive numerical studies demonstrate the advantages of the proposed method in terms of both estimation accuracy and computation efficiency. The method is also applied to the Worldwide Trading Networks and MovieLens 100K Dataset, yielding some interesting findings.

E0945: **Sparse Kronecker product decomposition: A general framework of signal region detection in image regression**
Presenter: **Long Feng**, University of Hong Kong, Hong Kong

The aim is to present the first Frequentist framework on signal region detection in high-resolution and high-order image regression problems. Image data and scalar-on-image regression have been intensively studied in recent years. However, most existing studies on such topics focused on outcome prediction, while the research on image region detection is rather limited, even though the latter is often more important. A general framework named Sparse Kronecker Product Decomposition (SKPD) is developed to tackle this issue. The SKPD framework is general in the sense that it works for both matrices and (high-order) tensors represented image data. This framework includes 1) the one-term SKPD, 2) the multiterm SKPD, and 3) the nonlinear SKPD. Nonconvex optimization problems are proposed to estimate the one-term and multiterm SKPDs, and path-following algorithms for nonconvex optimization are developed. The computed solutions of the path-following algorithm are guaranteed to converge to the truth with a particularly chosen initialization, even though the optimization is nonconvex. Moreover, the one-term and multiterm SKPD could also guarantee region detection consistency. The nonlinear SKPD is highly connected to shallow convolutional neural networks (CNN), particularly to CNN with one convolutional layer and one fully connected layer. Real brain imaging data in the UK Biobank database validate the effectiveness of SKPDs.

Tuesday 01.08.2023 13:20 - 15:00 Parallel Session C – EcoSta2023

EI003 Room 102 NEW DEVELOPMENTS OF BAYESIAN ECONOMETRICS AND STATISTICS Chair: Cathy W-S Chen

E0153: **Time-varying parameter heterogeneous autoregressive model with stochastic volatility**
Presenter: **Toshiaki Watanabe**, Hitotsubashi University, Japan
Co-authors: Jouchi Nakajima

The heterogeneous autoregressive (HAR) model performs well in volatility forecasting. This model formulates realized volatility (RV) as a function of past RVs with different frequencies, such as daily, weekly and monthly RVs. A method is proposed to extend the HAR model such that the coefficients of daily, weekly and monthly RVs and the error variance may change over time. The coefficients and the log of the error variance are assumed to follow first-order autoregressive processes. A Bayesian method using an efficient Markov chain Monte Carlo is developed to analyse the proposed model. An empirical application with the RV calculated using the 5-minute returns of the Nikkei 225 stock index is provided.

E0154: **The Bayesian lasso for variable selection of realized measures in a realized EGARCH model**
Presenter: **Richard Gerlach**, University of Sydney, Australia
Co-authors: Vica Tendenan, Chao Wang

The Realized EGARCH specification includes multiple realized measures in a financial time series volatility model. However, the question remains about which and how many realized measures to include in the model. The purpose is to employ the Lasso method in a Bayesian context to assist in selecting among a range of realized measures for inclusion. Markov chain Monte Carlo methods are designed for this purpose. We investigate the impacts of our approach on parameter estimation and forecasting of volatility and tail risk. Several competing models are employed for forecast comparison.

E0155: **Particle rolling MCMC with double-block sampling**
Presenter: **Yasuhiro Omori**, University of Tokyo, Japan
Co-authors: Naoki Awaya

An efficient particle Markov chain Monte Carlo methodology is proposed for the rolling-window estimation of state space models. The particles are updated to approximate the long sequence of posterior distributions as the estimation window is moved. To overcome the well-known weight degeneracy problem that causes the poor approximation, a practical double-block sampler with the conditional sequential Monte Carlo update is introduced, where one lineage from multiple candidates is chosen for the set of current state variables. The proposed sampler is justified in the augmented space through theoretical discussions. In the illustrative examples, it is shown to be successful in accurately estimating the posterior distributions of the model parameters.

EO044 Room 02 HIGH DIMENSIONAL DATA ANALYSIS, COPULA ESTIMATION, AND GENETIC STATISTICS Chair: Su-Yun Huang

E0463: **A parameterized empirical beta copula**
Presenter: **Xiaoling Dou**, Japan Womens University, Japan

The empirical beta copula is defined by using ranks of data and the beta distribution. It can be considered a non-parametric method for describing the dependence structure of multivariate data. Since no parameter is needed to estimate the copula, this method is very easy to use. However, since the method uses all ranks of the data, when the sample size is large, the computation of beta functions in this method shows difficulty. To enhance the utility of this copula, its parameterization is considered by separating the data into cells of a K by K grid and investigating the method of selecting K. This would make it easier to computation and provide a smoother copula density.

E0464: **An R package for QTL mapping and hotspot detection**
Presenter: **Chen-Hung Kao**, Institute of Statistical Science, Taiwan

Statistical methods for QTL mapping and QTL hotspot detection have been well-developed and used in various areas of biological studies. Here, an R package called QTLEMM is attempted to provide, that implements some commonly used and popular statistical methods of QTL mapping and QTL hotspot detection to explore the genetic architecture of quantitative traits, as well as the networks among QTL hotspots, genes and quantitative traits. An R package is provided for QTL mapping and hotspot detection, and a comprehensive overview of the primary functions with numerical and graphical outputs is described.

E0190: **On the efficiency-loss free ordering-robustness of product-PCA**
Presenter: **Hung Hung**, National Taiwan University, Taiwan
Co-authors: Su-Yun Huang

The robustness of the eigenvalue ordering, an important issue, is studied when estimating the leading eigen-subspace by principal component analysis (PCA). Previously, cross-data-matrix PCA (CDM-PCA) was proposed and shown to have smaller bias than PCA in estimating eigenvalues. While CDM-PCA has the potential to achieve a better estimation of the leading eigen-subspace than the usual PCA, its robustness is not well recognized. First, a more stable variant of CDM-PCA, is developed, called product-PCA (PPCA), that provides a more convenient formulation for theoretical investigation. Secondly, it is proved that, in the presence of outliers, PPCA is more robust than PCA in maintaining the correct ordering of leading eigenvalues. The robustness gain in PPCA comes from the random data partition, and it does not rely on a data down-weighting scheme as most robust statistical methods do. This enables us to establish the surprising finding that when there are no outliers, PPCA and PCA share the same asymptotic distribution. That is the robustness gain of PPCA in estimating the leading eigen-subspace has no efficiency loss in comparison with PCA. Simulation studies and a face data example are presented to show the merits of PPCA. In conclusion, PPCA has a good potential to replace the role of the usual PCA in real applications whether outliers are present or not.

E0408: **Contrastive principal component analysis in high dimension low sample size**
Presenter: **Shao-Hsuan Wang**, National Central University, Taiwan
Co-authors: Kazuyoshi Yata

Principal Component Analysis (PCA) is a commonly used linear dimensionality reduction method and is often used to visualize a single dataset; Contrastive Component Analysis (cPCA) can be used in situations where there are multiple datasets, and cPCA can explore the unique low-dimensional structure of a specific dataset on the premise of referring to other datasets. However, while cPCA has been shown in many fields to find important data patterns that PCA ignores, cPCA lacks a statistical model to identify why cPCA can identify those changes that are of interest. A statistical model for cPCA is proposed. The target data is divided into the signal matrix that is of interest and the nuisance matrix that is not of interest, and an effort is made to explain that cPCA can remove the influence of the nuisance matrix on the target data. On the other hand, the advantages of cPCA are illustrated in restoring the signal matrix using simulation analysis. Furthermore, a new method is proposed based on the model to help us decide on the contrast parameter that is important to perform cPCA. Finally, data patterns of interest in the synthetic image example are found by adjusting the contrast parameter and verifying that the new method of choosing the contrast parameter can achieve the same effect.

EO057 Room 03 THE STEIN METHOD, LIMIT THEOREMS AND APPLICATIONS Chair: Zhuosong Zhang

E0373: **BerryEsseen bounds for generalized U-statistics**
Presenter: **Zhuosong Zhang**, Southern University of Science and Technology, China

The focus is on optimal BerryEsseen bounds for the generalized U-statistics. The proof is based on a new BerryEsseen theorem for exchangeable pair approach by Stein's method under a general linearity condition setting. As applications, an optimal convergence rate of the normal approximation for subgraph counts in Erdos-Renyi graphs and graphon-random graphs is obtained.

E0561: **Non-asymptotic rates for random forest prediction intervals via Stein's method**
Presenter: **Krishnakumar Balasubramanian**, University of California, Davis, United States

Non-asymptotic rates for prediction intervals obtained using random forests using Stein's method will be presented. The previous work on viewing random forest predictions as adaptively weighted k-potential nearest neighbours prediction methods are leveraged to do so. This viewpoint is connected with recent advances in Stein's method for obtaining normal approximation bounds for region-stabilizing statistics to obtain our results for random forests. In particular, it is shown that k-potential nearest neighbours satisfy a certain region-stabilization property. Along the way, refined results on normal approximation were also obtained for a general class of region-stabilizing statistics, improving the recent results of other researchers.

E0693: **Asymptotic expansion of an estimator for the Hurst coefficient**
Presenter: **Hayate Yamagishi**, University of Tokyo, Japan
Co-authors: Nakahiro Yoshida, Yuliya Mishura

Asymptotic expansion is presented as an estimator of the Hurst coefficient of a fractional Brownian motion. First, the expansion formula of the principal term of the error of the estimator is derived using a recently developed theory of the asymptotic expansion of the distribution of Wiener functionals and utilizes the perturbation method on the obtained formula in order to calculate the expansion of the estimator. Numerical results show that the asymptotic expansion attains higher accuracy than the normal approximation.

E1065: **Cramer-type moderate deviation for quadratic forms with a fast rate**
Presenter: **Songhao Liu**, Southern University of Science and Technology, China
Co-authors: Xiao Fang, Qi-Man Shao

Let $X_1,\dots,X_n$ be independent and identically distributed random vectors in $\mathbb{R}^d$. Suppose $\mathbb{E}X_1 = 0$, $\mathrm{Cov}(X_1) = I_d$, where I_d is the $d \times d$ identity matrix. Suppose further that there exist positive constants t_0 and c_0 such that $\mathbb{E}e^{t_0|X_1|} \le c_0 < \infty$, where $|\cdot|$ denotes the Euclidean norm. Let $W = \frac{1}{\sqrt{n}}\sum_{i=1}^n X_i$ and let Z be a d-dimensional standard normal random vector. Let Q be a $d \times d$ symmetric positive definite matrix whose largest eigenvalue is 1. It is proved that for $0 \le x \le \varepsilon n^{1/6}$, $\left|\frac{\mathbb{P}(|Q^{1/2}W|>x)}{\mathbb{P}(|Q^{1/2}Z|>x)} - 1\right| \le C\left(\frac{1+x^5}{\det(Q^{1/2})n}1_{\{d\ge5\}} + \frac{1+x^3}{\det(Q^{1/2})n^{\frac{d}{d+1}}}1_{\{d\le4\}} + \frac{x^6}{n}\right)$, where ε and C are positive constants depending only on d, t_0, and c_0. This is the first extension of Cramér-type moderate deviation to the multivariate setting with a faster convergence rate than $1/\sqrt{n}$. The range of $x = o(n^{1/6})$ for the relative error to vanish and the dimension requirement $d \ge 5$ for the $1/n$ rate are both optimal. The result is proved using a new change of measure, a two-term Edgeworth expansion for the changed measure, and cancellation by symmetry for terms of the order $1/\sqrt{n}$.

EO072 Room Virtual R01 RECENT ADVANCES IN GAUSSIAN PROCESS THEORY AND APPLICATIONS Chair: Cheng Li

E0309: **Radial neighbors for provably accurate scalable approximations of Gaussian processes**
Presenter: **Yichen Zhu**, Duke University, United States

In geostatistical problems with massive sample sizes, Gaussian processes (GP) can be approximated using sparse-directed acyclic graphs to achieve scalable O(n) computational complexity. In these models, data at each location are typically assumed conditionally dependent on a small set of parents, which usually include a subset of the nearest neighbours. These methodologies often exhibit excellent empirical performance, but the lack of theoretical validation leads to unclear guidance in specifying the underlying graphical model. It may result in sensitivity to graph choice. These issues are addressed by introducing radial neighbours Gaussian processes and corresponding theoretical guarantees. The method proposes to approximate GPs using a sparse directed acyclic graph in which a directed edge connects every location to its neighbours within a predetermined radius. Using our novel construction, it is shown that one can accurately approximate a Gaussian process in Wasserstein-2 distance, with an error rate determined by the approximation radius, the spatial covariance function, and the spatial dispersion of samples. The method is also insensitive to specific graphical model choices. Further empirical validation of our approach is offered via applications on simulated and real-world data showing state-of-the-art performance in the posterior inference of spatial random effects.

E0318: **Maximum likelihood estimation for Gaussian processes under inequality constraints**
Presenter: **Francois Bachoc**, Universite Paul Sabatier, France

Covariance parameter estimation is considered for a Gaussian process under inequality constraints (boundedness, monotonicity or convexity) in fixed-domain asymptotic. The estimation of the variance parameter and the estimation of the micro ergodic parameter of the Matern and Wendland covariance functions are addressed. First, it is shown that the (unconstrained) maximum likelihood estimator has the same asymptotic distribution, unconditionally and conditionally, to the fact that the Gaussian process satisfies the inequality constraints. Then, the recently suggested constrained maximum likelihood estimator is studied. It is shown that it has the same asymptotic distribution as the (unconstrained) maximum likelihood estimator. In addition, it is shown in simulations that the constrained maximum likelihood estimator is generally more accurate on finite samples.

E0427: **Inferring manifolds from noisy data using Gaussian processes**
Presenter: **Nan Wu**, University of Texas at Dallas, United States
Co-authors: David Dunson

The focus is on the study of a noisy data set sampled around an unknown Riemannian submanifold of a high-dimensional space. Most existing manifold learning algorithms replace the original data with lower dimensional coordinates without providing an estimate of the manifold in the observation space or using the manifold to denoise the original data. A Manifold reconstruction is proposed via the Gaussian processes (MrGap) algorithm for addressing these problems, allowing interpolation of the estimated manifold between fitted data points. The proposed approach is motivated by novel theoretical properties of local covariance matrices constructed from noisy samples on a manifold. The results enable us to turn a global manifold reconstruction problem into a local regression problem, allowing the application of Gaussian processes for probabilistic manifold reconstruction. The classical manifold learning algorithms are reviewed, and the theoretical foundation of the new method, MrGap, is discussed. Simulated and real data examples will be provided to illustrate the performance.

E0500: **Distributed Bayesian varying coefficient modeling using a Gaussian process prior**
Presenter: **Sanvesh Srivastava**, The University of Iowa, United States

The divide-and-conquer technique is used to address inefficient inference in Varying coefficient models (VCMs) based on Gaussian process (GP) priors. Our proposal has three steps. The first step creates many data subsets with much smaller sample sizes by sampling without replacement from the full data. The second step formulates VCM as a linear mixed-effects model and develops a data augmentation (DA)-type algorithm for obtaining MCMC draws of the parameters and predictions on all the subsets in parallel. The DA-type algorithm appropriately modifies the likelihood such that every subset posterior distribution accurately approximates the corresponding true posterior distribution. The third step develops a combination algorithm for aggregating MCMC-based estimates of the subset posterior distributions into a single posterior distribution called the Aggregated Monte Carlo (AMC) posterior. The AMC posterior has minimax optimal posterior convergence rates in estimating the varying coefficients and the mean regression function.

EO225 Room 503 ADVANCES IN BAYESIAN NONPARAMETRICS AND MODEL-BASED CLUSTERING Chair: Beatrice Franzolini

E0630: **Adaptive prior distributions for record linkage tasks**
Presenter: **Brenda Betancourt**, NORC at the University of Chicago, United States

In database management, record linkage aims to identify multiple records that correspond to the same individual. Record linkage can be treated as a clustering problem in which one or more noisy database records are associated with a unique latent entity. In contrast to traditional clustering applications, a large number of clusters with a few observations per cluster is expected. Hence, two new classes of prior distributions based on exchangeable sequences of clusters and allelic partitions are proposed for the small cluster setting of record linkage. The proposed priors facilitate the introduction of information about the cluster size distribution at different scales and naturally enforce sublinear growth of the maximum cluster size, known as the micro clustering property. In addition, a set of novel micro clustering conditions are introduced to impose further constraints on the cluster sizes a priori. The performance of the proposed classes of priors is evaluated using simulated data and official statistics data sets.

E0680: **Partially exchangeable stochastic block models for multilayer networks**
Presenter: **Francesco Gaffi**, University of Notre Dame, United States
Co-authors: Daniele Durante, Antonio Lijoi, Igor Pruenster

There is an increasing availability of multilayer network data but still a lack of state-of-the-art models for node-coloured multilayer networks, which can flexibly account for both within and across-layer block-connectivity structures while incorporating layer information in a principled probabilistic manner. Such a gap is covered by proposing a new class of partially exchangeable stochastic block models that relies on a hierarchical random partition prior to the group allocations of nodes driven by the urn scheme of a hierarchical normalized completely random measure. The partial exchangeability assumption among nodes according to layer partitions allows inferring both within- and across-layer blocks while preserving probabilistic coherence, principled uncertainty quantification and formal inclusion of prior information from the layer division. The mathematical tractability and projectivity of the construction further allow analytically deriving predictive within- and across-layer co-clustering probabilities, thus facilitating prior elicitation and development of rigorous predictive strategies for both the connections and allocations of future incoming nodes. The practical performance of this novel class is illustrated in simulation studies and in a real-world criminal network application, where the proposed model displays clear gains relative to alternative solutions in estimation, uncertainty quantification and prediction.

E0360: **Measuring the impact of the prior in Bayesian nonparametrics via optimal transport**
Presenter: **Marta Catalano**, University of Warwick, United Kingdom

The Dirichlet process has been pivotal to the development of Bayesian nonparametric, allowing one to learn the law of the observations through closed-form expressions. Still, its learning mechanism is often too simplistic, and many generalizations have been proposed to increase its flexibility, a popular one being the class of normalized completely random measures. A simple yet fundamental matter is investigated: will a different prior actually guarantee a different learning outcome? To this end, a new framework is developed for assessing the merging rate of opinions based on three leading pillars: i) the investigation of identifiability of completely random measures; ii) the measurement of their discrepancy through a novel optimal transport distance; iii) the establishment of general techniques to conduct posterior analyses, unravelling both finite-sample and asymptotic behaviour of the distance as the number of observations grows. The findings provide neat and interpretable insights on the impact of popular Bayesian nonparametric priors, with very mild assumptions on the data-generating process.

E0433: **Bayesian clustering of high-dimensional data via latent repulsive mixtures**
Presenter: **Alessandra Guglielmi**, Politecnico di Milano, Italy

Model-based clustering of moderate or large dimensional data is notoriously difficult. A model is proposed for simultaneous dimensionality reduction and clustering by assuming a mixture model for a set of latent scores, which are then linked to the observations via a Gaussian latent factor model. Other researchers recently investigated this approach. A factor-analytic representation is used, and a mixture model is assumed for the latent factors. However, performance can deteriorate in the presence of model misspecification. Assuming a repulsive point process prior to the component-specific means of the mixture for the latent scores is shown to yield a more robust model that outperforms the standard mixture model for the latent factors in several simulated scenarios. To favour well-separated clusters of data, the repulsive point process must be anisotropic, and its density should be tractable for efficient posterior inference. These issues are addressed by proposing a general construction for anisotropic determinantal point processes.

E0505: **Repulsion, chaos and equilibrium in mixture models**
Presenter: **Andrea Cremaschi**, ASTAR, Singapore
Co-authors: Maria De Iorio, Timothy Wertz

Mixture models are commonly used to analyse data presenting heterogeneity and overdispersion, as they allow the identification of subpopulations. In the Bayesian framework, this entails the specification of suitable prior distributions for the weights and location parameters of the mixture. Widely used are Bayesian semi-parametric models based on mixtures with infinite or random number of components. Often, the flexibility of these models does not translate into interpretability of the identified clusters. To overcome this issue, clustering methods based on repulsive mixtures have been recently proposed, including a repulsive term in the prior distribution of the atoms of the mixture, favouring locations far apart. This approach is increasingly popular and allows to produce well-separated clusters, thus facilitating the interpretation of the results. However, the resulting models are usually not easy to handle due to the introduction of unknown normalising constants. Exploiting results from statistical mechanics, a novel class of repulsive prior distributions is proposed based on Gibbs measures associated with joint distributions of eigenvalues of random matrices, which naturally possess a repulsive property. The proposed framework greatly simplifies the computations needed due to the availability of the normalising constant in closed form. The novel class of priors and their properties is illustrated as well as their clustering performance, on benchmark datasets.

EO195 Room 506 RECENT ADVANCES IN STATISTICAL MODELING Chair: Eftychia Solea

E0283: **Poisson PCA for matrix count data**
Presenter: **Joni Virta**, University of Turku, Finland
Co-authors: Andreas Artemiou

A dimension reduction framework is developed for data consisting of matrices of counts. Our model is based on the assumption of the existence of a small amount of independent normal latent variables that drive the dependency structure of the observed data and can be seen as the exact discrete analogue of a contaminated low-rank matrix normal model. Estimators are derived for the model parameters, and their limiting normality is established. An extension of a recent proposal from the literature is used to estimate the latent dimension of the model. The method is shown to outperform both its vectorization-based competitors and matrix methods, assuming the continuity of the data distribution in analysing simulated data and real-world abundance data.

E0457: **On sufficient graphical model**
Presenter: **Kyongwon Kim**, Ewha Womans University, Korea, South

A sufficient graphical model is introduced by applying the recently developed nonlinear sufficient dimension reduction techniques to evaluate conditional independence. The graphical model is nonparametric in nature, as it does not make distributional assumptions such as the Gaussian or copula Gaussian assumptions. However, unlike a fully nonparametric graphical model, which relies on the high-dimensional kernel to characterize conditional independence, the graphical model is based on conditional independence, given a set of sufficient predictors with a substantially reduced dimension. In this way, the curse of dimensionality that comes with a high-dimensional kernel is avoided. The estimate's population-level properties, convergence rate, and variable selection consistency are developed. Simulation comparisons and an analysis of the DREAM 4 Challenge data set demonstrate that the method outperforms the existing methods when the Gaussian or copula Gaussian assumptions are violated. Its performance remains excellent in the high-dimensional setting.

E0598: **A concave pairwise fusion approach to multiresponse regression clustering**
Presenter: **Abdul-Nasah Soale**, Case Western Reserve University, United States
Co-authors: Yuexiao Dong, Chen Chen

Classical multiresponse regression studies the effect of a set of predictors on multiple responses. The problem is considered that the samples from the regression model consist of subgroups with different mean values and the same predictor effects. A concave penalized regression approach is used to detect the subgroups or clusters by shrinking the pairwise differences of the mean values. The proposed method exploits heterogeneity and automatically divides the observations into subgroups. The new method's performance is compared to the univariate response regression clustering method in simulation studies. An analysis of healthy and Parkinson's disease patients is also provided.

E0808: **Nonparametric estimation for IID paths of SDE perturbed by Levy noise**
Presenter: **Subhra Sankar Dhar**, IIT Kanpur, India
Co-authors: Debopriya Mukharjee

The focus is on the problem of nonparametric estimators of the drift and diffusion coefficients of the trend for stochastic differential equations driven by Levy noise. The methodology adopted is fully nonparametric, and the coefficients are computed from independent continuous observations on a compact time interval. Almost sure uniform asymptotic convergence rates of the proposed estimators are achieved as the number of observed curves moves to infinity.

EO074 Room 603 FISCAL AND MONETARY POLICIES Chair: Etsuro Shioji

E0661: **Central bank information effects in Japan: The role of uncertainty channel**
Presenter: **Hiroshi Morita**, Tokyo Institute of Technology, Japan
Co-authors: Taiki Ono, Ryo Matsumoto

Central bank information effects have been analyzed in the recent literature on monetary policy. A recent identification method is applied to the Japanese data to empirically examine the macroeconomic effects of central bank information shock and pure monetary policy shock. These shocks are identified by combining high-frequency identification and sign restriction. The empirical results support the presence of central bank information effects in Japan. Particularly, the central bank information shock accompanying monetary tightening decreases economic uncertainty and increases stock prices and output, suggesting that the central bank's optimistic outlook is conveyed through contractionary monetary actions. The results of the forecast error variance decomposition indicate that the central bank's information effect may be spread through changes in uncertainty. Finally, the total effect of monetary policy and information shocks on the variables is much larger than that of the shocks identified by the conventional Cholesky decomposition. These findings are important for evaluating the true effects of monetary actions on the economy.

E0371: **The signaling effects of fiscal announcements**
Presenter: **Anna Rogantini Picco**, Sveriges Riksbank, Sweden
Co-authors: Leonardo Melosi, Francesco Zanetti, Hiroshi Morita

Fiscal announcements may transfer information about the government's view of the macroeconomic outlook to the private sector, diminishing the effectiveness of fiscal policy as a stabilization tool. A simple model is developed that transparently outlines conditions and key properties of the signalling effect and guides the empirical tests, and it is shown that results hold in a standard micro-founded model. A novel dataset that combines daily data on Japanese stock prices is constructed with narrative records from press releases about a set of extraordinary fiscal packages introduced by the Japanese government from 2011-2020. These fiscal stimuli are shown that be often interpreted as negative news by the stock market, whereas exogenous fiscal interventions that do not convey any information about the business cycle (e.g., the successful bids to host the Olympics on September 8, 2013) fostered bullish reactions. In addition, these negative effects on stock prices arose more commonly when fiscal stimuli were announced against a backdrop of heightened macroeconomic uncertainty. Our empirical findings support the theory of signalling effects.

E0684: **Dark matter of Japanese government bonds**
Presenter: **Toshitaka Sekine**, Hitotsubashi University, Japan

Why has the long-term Japanese Government Bonds (JGB) yield remained so low despite the extremely high debt outstanding? Is there dark matter that the JGB yield has gravitated to? An attempt was made to quantify the following two effects using the Demand System Approach based on our age-decomposition of net financial positions of the household sector. One is the effect of unconventional monetary policy by the Bank of Japan (BoJ). As BoJ has purchased JGBs so much, there is less supply of JGBs in the market, and the average maturity of JGBs has become shorter. The other is the effect of household savings. Against the risks of health and longevity, elderly families have, in effect, increased their possession of JGBs through financial intermediaries, as JGBs are special assets in terms of liquidity and safety.

E0496: **Yield curve control under attack: Where do the pressures come from?**
Presenter: **Etsuro Shioji**, Hitotsubashi University, Japan

In recent years, central banks worldwide have adopted various types of unconventional monetary policies. Since 2016, the Bank of Japan has implemented the Yield Curve Control Policy and set a target range for the yields on 10-year government bonds. A potential shortcoming of such a policy is that it could invite speculative attacks from investors, especially if the bank's commitment to defend the "red line" is deemed not fully credible. A new measure of market pressures exerted on the upper and the lower limits specified by the central bank is constructed. Like the Exchange Market Pressure Index, which is widely used in the international finance literature, the proposed index combines information on the movement of the bond yields with the amount of market intervention made by the bank. On the other hand, it is considered the target-zone-like feature of the YCC in the index specification. The determinants of this novel "Bond Market Pressure Index" are studied, and the relative importance of external vs internal factors is compared. It is found that the external factor, represented by the US long-term interest rate, plays a much more important role than the domestic factor, represented by the Japanese CPI.

EO196 Room 604 RECENT ADVANCES IN METHODOLOGICAL ECONOMETRICS Chair: Chu-An Liu

E0685: **Long memory and structural breaks in testing the implied-realized volatility relation**
Presenter: **Yi-Chi Chen**, National Cheng Kung University, Taiwan

The recent literature has shown that the implied-realized volatility relation can be well-modelled as fractional integration and thus provide evidence for the implied volatility unbiasedness. On the other hand, it has been found that the observed long-range dependence could be generated by structural breaks, in particular, uncommon breaks. The unbiasedness hypothesis is revisited, and the problem is reformulated as testing for cross-correlation between two volatility series. This testing procedure takes advantage of the residual cross-correlation function by first pre-whitening two-time series. Such a correlation test is free from parametric volatility models that depend on specific distributional assumptions and are subject to size distortions. Simulation studies will be considered to demonstrate the finite sample properties of the proposed correlation test as opposed to its conventional counterparts. The volatility dynamics for the major exchange rates are investigated to empirically verify whether implied volatility is an unbiased forecast of realized volatility and whether the unbiasedness anomaly remains a statistical artefact in a pairwise cross-dependence testing framework.

E0305: **Model averaging prediction for possibly nonstationary autoregressions**
Presenter: **Chu-An Liu**, Academia Sinica, Taiwan
Co-authors: Tzu-Chi Lin

As an alternative to model selection (MS), model averaging (MA) is considered for integrated autoregressive processes of infinite order. A uniformly asymptotic expression for the mean squared prediction error (MSPE) of the averaging prediction with fixed weights and then propose a Mallows-type criterion to select the data-driven weights that minimize the MSPE asymptotically. The proposed MA estimator and its variants, Shibata and Akaike MA estimators, are asymptotically optimal in the sense of achieving the lowest possible MSPE. further. MA can provide significant MSPE reduction over MS when the model misspecification bias is algebraic decay. These theoretical findings are supported by Monte Carlo simulations and real data analysis.

E0613: **Robust estimation of the quantile mediation treatment effect with machine learning**
Presenter: **Yu-Min Yen**, National Chengchi University, Taiwan
Co-authors: Martin Huber, Yu-Chin Hsu

The quantile mediation treatment effect estimation is studied using a double/debias estimator. The nuisance parameters of the proposed estimator are estimated with a machine-learning method, and cross-fitting is used to reduce estimation bias from overfitting and/or regularization of the machine learner. A multiplier bootstrap procedure is then used to conduct statistical inferences. Relevant uniform consistency of the proposed estimator and uniform validity of the multiplier bootstrap procedure is established. The performance of the proposed estimator is illustrated by conducting a simulation. Then a sensitivity analysis for the proposed estimator and a possible extension to allow for sequential mediators are discussed.

E0453: **Nonlinear least squares, model selection, and model averaging for social interaction models**
Presenter: **Hon Ho Kwok**, National Taiwan University, Taiwan

A unified theory of nonlinear least squares, model selection, and model averaging is developed for social interaction models. First, although the generalized method of moments and maximum likelihood have been used to estimate social interaction models, nonlinear least squares consistently can also consistently estimate social interaction models. Since the least squares criterion functions coincide with Mallows' criterion, a model selection and averaging theory based on Mallows' criterion is provided. Second, an identification theory is developed based on the least squares criterion. Third, Mallows' criterion is proposed to be used for selecting networks, variables, and models.

EO086 Room 605 STATISTICAL THEORY FOR STOCHASTIC PROCESSES Chair: Teppei Ogihara

E0389: **Asymptotically efficient estimation for diffusion processes with nonsynchronous observations**
Presenter: **Teppei Ogihara**, University of Tokyo, Japan

The properties of a maximum-likelihood-type estimation method are investigated for nonsynchronous observations of two-dimensional diffusion processes. Nonsynchronous observation is a fundamental issue in high-frequency financial data, where the observation time of stock prices may not necessarily coincide due to trades occurring at different times. The limit where the observation time diverges to infinity is considered, and asymptotic properties are demonstrated as consistency and asymptotic normality of the estimator. Moreover, the asymptotic efficiency of the proposed method is discussed.

E0526: **Quasi-likelihood analysis for Student-Levy regression**
Presenter: **Yuma Uehara**, Kansai University, Japan
Co-authors: Lorenzo Mercuri, Hiroki Masuda

A linear regression model driven by a Student Levy process with constant scale and arbitrary degrees of freedom is considered. It is supposed that the data is observed at high frequency over a long period. In the proposed model, three estimation targets are considered: trend, scale and the degrees of freedom in the driving noise. However, the distribution of the Student-Levy process at a small time t is not given under closed form, and this makes it difficult to estimate the above three unknown quantities directly from the high-frequency data. Hence the following stepwise estimation procedure is introduced for our model. First, the trend and scale parameter is estimated by Cauchy quasi-likelihood which is based on the fact that the (appropriately scaled) distribution of the Student-Levy process tends to Cauchy distribution as the time tends to 0. After that, the rest parameter by constructing unit-time residual and maximum likelihood estimation is estimated. Also, its asymptotic behaviour, and its simulation aspects are presented.

E0610: **Asymptotically efficient estimation for mixed fractional Brownian motion under high-frequency observations**
Presenter: **Tetsuya Takabatake**, Hiroshima University, Japan

The focus is on an asymptotically efficient estimation for a mixed fractional Brownian under high-frequency observations. Namely, a local asymptotic normality property for this setup is shown, and an asymptotically efficient estimator based on a one-step estimator using a quadratic variation-type estimator with pre-averaged data as an initial guess is proposed.

E0806: **Estimating Hawkes processes from observations of a sample path at discrete times points**
Presenter: **Feng Chen**, UNSW Syd, Australia

Estimation of the Hawkes process from complete observations of a sample path is relatively straightforward using either the maximum likelihood or other methods. However, estimating the parameters of a Hawkes process from observations of a sample path at discrete time points only is challenging due to the intractability of the likelihood with such data. A method is introduced to estimate the Hawkes process from a discretely observed sample path. The work relies on a state-space representation of the problem and a sequence Monte Carlo (aka particle filtering) approximation to the likelihood function. The performance of the proposed method is evaluated using simulation experiments and it is compared with some recently published benchmark methods in the literature.

EO075 Room 606 DEVELOPMENTS IN TIME SERIES MODELLING AND INFERENCE Chair: Feiyu Jiang

E0437: **Transformed cointegration models with partially linear additivity**
Presenter: **Yingqian Lin**, Shanghai University of Finance and Economics, China
Co-authors: Yundong Tu

Two general classes of transformed cointegration models are considered. In the first class of models, the dependent variable, after a parametric monotonic transformation, is cointegrated with nonlinear parametric functions of unit root regressors and stationary regressors in a linear form. The second class augment the first class with unknown integrable functions of the unit root regressors in an additive way. Extremum estimators for the parameters in the transformation function, plug-in estimators for parameters in the linear components, and sieve estimators for the unknown functions are presented. Asymptotic properties of the proposed estimators are developed, which are shown to depend on the transformation, the functions and model parameters. The theory is further extended to allow for both endogeneity of the nonstationary regressors and serially dependent errors. Numerical results demonstrate the nice performance of the estimators, corroborate the theoretical development and illustrate the practical merits of the proposed models.

E0627: **Multistep forecast averaging with stochastic and deterministic trends**
Presenter: **Xuewen Yu**, Fudan University, China
Co-authors: Mohitosh Kejriwal, Linh Nguyen

A new approach to constructing multistep combination forecasts in a nonstationary framework with stochastic and deterministic trends is presented. Existing forecast combination approaches in the stationary setup typically target the in-sample asymptotic mean squared error (AMSE) relying on its approximate equivalence with the asymptotic forecast risk (AFR). Such equivalence, however, breaks down in a nonstationary setup. Combination forecasts are developed based on minimizing an Accumulated Prediction Errors (APE) criterion that directly targets the AFR and remains valid whether the time series is stationary or not. It is shown that the performance of APE-weighted forecasts is close to that of the optimal, infeasible combination forecasts. Monte Carlo experiments are used to demonstrate the finite sample efficacy of the proposed procedure relative to Mallows/Cross-Validation weighting that targets the AMSE. An application to forecasting US macroeconomic time series demonstrates the relevance of the proposed method in practice.

E0510: **Matrix GARCH model: Inference and applications**
Presenter: **Cheng Yu**, Tsinghua University, China
Co-authors: Ke Zhu, Feiyu Jiang, Dong Li

Matrix-variate time series data are largely available in plenty of applications. However, no attempt has been made to study their conditional heteroskedasticity, which is often observed in economic and financial data. To fill the gap, a new matrix generalized autoregressive conditional heteroskedasticity (GARCH) model is proposed, which can capture the dynamics of conditional row and column covariance matrices of the matrix time series. The key element of the matrix GARCH model is a univariate GARCH-type specification for the trace of conditional row or column covariance matrix since the conditional row, and column covariance matrices can not be identified without this trace specification. Moreover, the quasi-maximum likelihood estimator is proposed for the model, and the portmanteau tests are constructed for model diagnostic checking. To handle the large dimensional matrix time series, a matrix factor GARCH model is further raised. Finally, three applications are given on credit default swap prices, global stock sector indices, and future prices to demonstrate the advantage of the matrix GARCH and matrix factor GARCH models over the existing multivariate GARCH-type models in risk management and portfolio allocation.

E0710: **Estimation based on martingale difference divergence**
Presenter: **Kunyang Song**, The University of Hong Kong, Hong Kong

Finding valid instrumental variables (IVs) is important but hard in the linear regression model. However, the classical estimation method, such as the 2-stage least square estimator, is not applicable when the linear regression model is under-identified without enough valid IVs. Based on the martingale difference divergence (MDD), a new estimator is proposed for the general nonlinear regression model, and this estimator is applicable even when the regression model is under-identified. Under certain regular conditions, the consistency and asymptotic normality of this MDD-based estimator is established. As an extension, a new MDD-based loss function with an additional L_1 penalty is proposed to select non-zero parameters in several kinds of deep neural networks and further enables the causal discovery. Simulations are also given to illustrate the importance of the proposed estimators.

EO111 Room 701 NETWORK DATA ANALYSIS Chair: Frederick Kin Hing Phoa

E0755: **Inferring associations along the causal chains in a network**
Presenter: **Chen-Hsiang Yeang**, Academia Sinica, Taiwan

In a large system where the relations between entities are represented as a network, the effects of perturbing entities are propagated along the paths in the network. Many perturbations and responses may occur concurrently in the same experiments or observational events. Therefore, it is generally difficult to distinguish the causal relations from the spurious associations in such a system. Moreover, although the causal effects can, in principle, propagate along all directions permitted by the network structure, in reality, the majority of the causal effects are likely mediated by the paths restricted to a compact subnetwork. Finding this core subnetwork from the massive amount of association data also remains an open problem. Several algorithms are proposed to tackle these two open problems. First, a model selection procedure is built, which prioritizes the putative associations by their path lengths connecting the perturbations and effects and iteratively incorporates the candidate associations that best fit the data. Second, a network diffusion model is constructed, which assigns edge weights according to the paths traversing perturbations-effects pairs, and develops an algorithm to extract the core subnetwork with high edge weights. These two algorithms are applied to the multi-omics data of 33 cancer types in The Cancer Genome Atlas (TCGA) and identify the Integrated Hierarchical Association Structure (IHAS) within and across the cancer types.

E0819: **Link prediction by exploring common neighborhoods**
Presenter: **Tso-Jung Yen**, Academia Sinica, Taiwan

Social network analysis aims to establish the properties of a network by exploring the link structure of the network. However, due to concerns such as confidentiality and privacy, a social network may not provide full information on its links. As some of the links are missing, it is difficult to establish the network's properties by exploring its link structure. A method for recovering such missing links is proposed. It is paid attention to a situation in which some nodes have fully-observed links. The method relies on exploiting the sub-network of these "anchor" nodes to recover missing links of nodes that have neighbourhoods overlapping with the "anchor" nodes. It uses a graph neural network to extract information from these neighbourhoods and then applies this information to a regression model for missing link recovery. This method is demonstrated on real-world social network data. The results show that this method can achieve better performance than traditional methods that are solely based on node attributes for missing link recovery.

E0826: **Quantifying biodiversity from network perspectives**
Presenter: **Wei-chung Liu**, Academia Sinica, Taiwan

Biodiversity traditionally concerns the number of species and their genomic diversity in an ecosystem. Since species are linked trophically in an ecosystem, it is argued that biodiversity can also be viewed from a network perspective regarding the diversity of their interaction patterns. Three approaches are proposed. The first is based on the notion of regular equivalence, measuring the similarity between species positions in an ecological network; the second is based on the concept of positional centrality, where diversity is defined in terms of how diverse species centrality values are in an ecological network. Third is based on the ecological concept of direct and indirect interactions, and diversity here extends from how similarly two species affect all individual species in the same network. Then the linkage is investigated between those three biodiversity measures, various structural properties of an ecological network, and some species traits. The discussion then focuses on the future development of this particular field of network research in ecology.

E1010: **A uniform placement of alters on spherical surface for ego-centric network with community structure and alter attributes**
Presenter: **Chao-Hui Huang**, National Tsing Hua University, Taiwan
Co-authors: Frederick Kin Hing Phoa

An ego-centric network describes the relationships between a particular node (ego) to its neighbouring nodes (alters), so it is essential to present such a network with good visualization. The aim is to introduce an efficient method, namely the Uniform Placement of Alters on a Spherical Surface (U-PASS), to represent an ego-centric network so that all alters are scattered uniformly on the surface of the unit sphere. Unlike other simple uniformity that considers maximizing Euclidean distances among nodes, U-PASS is a three-stage method that spreads the alters considering existing edges among alters, no overlapping of node clusters, and node attribute information. Particle swarm optimization is employed to improve efficiency in node allocations. The connection between the U-PASS and the minimum energy design on a two-dimensional flat plane with a specific gradient is shown to guarantee uniformity. A demonstration is provided on allocating nodes of an ego-centric network with 50 nodes, and some distance statistics show good performance of U-PASS when compared to four state-of-the-art methods via self-organizing maps and force-driven approaches.

EO064 Room 702 STATISTICAL MODELS FOR COMPLEX BRAIN IMAGING DATA Chair: Shuo Chen

E0253: **Covariance outcome modelling via covariate assisted principal regression**
Presenter: **Xi Luo**, Univ of Texas Health Science Center at Houston, United States
Co-authors: Yi Zhao, Brian Caffo, Bingkai Wang

Modelling the variations in covariance matrix outcomes is becoming an important topic in many fields, including financial and neuroimaging analysis. The problem of regressing covariance matrices on vector covariates collected from each observational unit in cross-sectional or longitudinal settings is considered. The aim is to review the proposed (generalized) linear model framework and recent advances for this problem. The focus will be on mathematical formulation, algorithmic development, and asymptotic properties. Accuracy and robustness will be demonstrated using extensive simulations. The proposals were also applied to a few large-scale resting-state functional magnetic resonance imaging studies, and the specific human brain network changes associated with covariates were identified.

E0312: **Regression frameworks for brain network distance metrics**
Presenter: **Sean Simpson**, Wake Forest University School of Medicine, United States

Brain network analyses have exploded in recent years and hold great potential in helping us understand normal and abnormal brain function. Network science approaches have facilitated these analyses and understanding of how the brain is structurally and functionally organized. However, the development of statistical methods relating this organization to health outcomes has lagged behind. An attempt has been made to address this need by developing regression frameworks for brain network distance metrics that allow relating system-level properties of brain networks to outcomes of interest. These frameworks serve as synergistic fusions of statistical approaches with network science methods, providing needed analytic foundations for whole-brain network data. These approaches developed for single-task, multi-task/multi-session, and multilevel brain network data are delineated. These tools help expand the suite of analytical tools for whole-brain networks and aid in providing complementary insight into brain function.

E0672: **Statistical modeling for positronium lifetimeimage reconstruction using time-of-flight positron emission tomography**
Presenter: **Hsin-Hsiung Huang**, University of Central Florida, United States

Positron emission tomography (PET) has been widely used to diagnose serious diseases, including cancer and Alzheimer's disease, based on the uptake of radiolabeled molecules that target certain pathological signatures. Recently, a novel imaging mode known as positronium lifetime imaging (PLI) has also been shown to be possible with time-of-flight (TOF) PET. PLI is also of practical interest because it can provide complementary disease information reflecting conditions of the tissue microenvironment via mechanisms that are independent of tracer uptake. However, for the present practical systems with a finite TOF resolution, the PLI reconstruction problem has yet to be fully formulated to develop accurate reconstruction algorithms. This paper addresses this challenge by developing a statistical model for the PLI data and deriving from it a maximum-likelihood algorithm for reconstructing lifetime images alongside the uptake images. Using realistic computer simulation data shows that the proposed algorithm can produce quantitatively accurate lifetime images for a 570 ps TOF PET system. The recent findings about the parameter effects for the reconstruction bias and variance analysis and the real-world PLI image with mixed exponential distributions are also presented.

E0859: **Graph neural network for fMRI and EEG brain data analysis**
Presenter: **Don Hong**, Middle Tennessee State University, United States

Functional connections in the brain have provided significant information to explain various pathological conditions and behavioural characteristics. The use of a graph neural network called graphSAGE to investigate resting-state fMRI and the applications of machine learning-based multi-modal approaches to combine EEG measures with fMRI to observe neurophysiological events in high temporal and spatial resolution will be discussed.

EO056 Room 703 RECENT DEVELOPMENTS IN DESIGN OF EXPERIMENTS Chair: Jian-Feng Yang

E0191: **Group-orthogonal subsampling for big data linear mixed models**

Presenter: **Fasheng Sun**, Northeast Normal University, China

The linear mixed model is a popular and common modelling method in statistical analysis. It is computationally challenging to obtain parameter estimates for big data in the linear mixed model. The current subsampling methods are aimed at the situation where the data is independent without considering the correlation within the data. An optimal subsampling method for a linear mixed model based on maximizing the determinant of the variance-covariance matrix of the subsampling estimator is proposed. The proposed subsampling procedure is also optimal under the A-optimality criterion, which minimizes the trace of the variance-covariance matrix of the subsampling estimator. Furthermore, the asymptotic property of the subsampling estimator is established. Numerical examples based on both simulated and real data are provided to illustrate the proposed subsampling method.

E0192: **Efficient Kriging using designs with low fill distance and high separation distance**

Presenter: **Xu He**, Chinese Academy of Sciences, China

Kriging is a powerful technique to emulate computer experiments. Accurate Kriging requires good experimental designs. The purpose is to construct space-filling designs that are most suitable for Kriging. Firstly, the study provides evidence that although it is well-known that designs with low fill distance or high separation distance are appealing for Kriging, the two distance measures should be combined to reach a more striking criterion for selecting designs. Secondly, a method is proposed to construct optimal designs under the newly proposed criterion efficiently. Thirdly, a robust method is provided to transform design points towards boundary facets of the input space. Numerical results suggest that the new criterion, construction, and transformation combined perform robustly well in Kriging interpolation under various scenarios.

E0195: **Schematic array and its modification**

Presenter: **Yu Tang**, Soochow University, China

The association scheme was first introduced by statisticians in connection with the design of experiments and has been proven very useful in many fields, including permutation groups, graphs, and coding theory. An array is called a schematic if it runs from an association scheme concerning distance. Schematic arrays, especially schematic orthogonal arrays, have been ideal tools for designing experiments and generating software test suites. However, the definition of the original schematic array is too demanding. It requires the relationship between any two distinct rows and overemphasizes the single-row property. This drawback dramatically limits the existence of results on schematic arrays. The original conditions of the association scheme are modified, and the concept of the revised schematic array is proposed. The rationality of the revised definition is further elaborated. Finally, Two examples of revised schematic arrays are also provided, including three-level mirror-symmetric orthogonal arrays and Latin hypercube designs.

E0206: **The order-of-addition experiments on the adjacency relationship**

Presenter: **Jian-Feng Yang**, Nankai University, China

With the development of science and technology, the scale of data has become huge. The design of experiments plays a more and more critical role in different areas. In many cases, the addition order of components, which is the order-of-addition (OofA) problem, is studied. The adjacency relationship (AR) between components usually has an important impact on the final results. However, there is still very little research on this part, while most of the existing order-of-addition experiments focus on the relative or absolute positions between components. How AR between components affects the response is considered. Starting from the Traveling Salesman Problem, the aim is to propose an AR model and presents an algorithm for inferring the optimal order. In addition, a loop-full design, which is different from the OofA full design, is proposed. The properties of the loop-full design and the D-optimality under the AR model are given. A kind of minimal-point design is also put forward, which greatly reduces the run size and performs well in designs with the same run size. Based on the minimal-point design, the study brings up an improved algorithm to improve D-efficiency. The simulation results show the effectiveness of the model and design.

EO121 Room 704 SKEW DISTRIBUTIONS ON THE CIRCLE AND THEIR APPLICATIONS Chair: Tomoaki Imoto

E0450: **A cylindrical hidden Markov model based on skewed circular distributions**

Presenter: **Yoichi Miyata**, Takasaki City University of Economics, Japan

Co-authors: Takayuki Shiohama, Toshihiro Abe

Hidden Markov models are known to be a useful method for estimating the timing of structural change and the structure of each population for time series data. Cylindrical hidden Markov models are applied when data consist of pairs of non-negative values and values on the circle. If angular data show somewhat asymmetric patterns, it is required to model a circular marginal distribution to have a slightly more strongly skewed shape. A hidden Markov model whose components are cylindrical distributions in which a circular random variable (representing the angle) has an extended sine-skewed circular distribution. In particular, some conditions are clarified for the consistency of the maximum likelihood estimator and present numerical examples of the model applied to real data.

E0504: **Data-based bin width selection for rose diagram**

Presenter: **Yasuhito Tsuruta**, The University of Nagano, Japan

A rose diagram is a representation that circularly organizes data with the bin width as the central angle. This diagram is widely used to display and summarize circular data. Some studies have proposed the selector of bin width based on data. However, only a few papers have discussed the property of these selectors from a statistical perspective. Thus, the aim is to provide a data-based bin width selector for rose diagrams using a statistical approach such as minimizing an error criterion. The radius of the rose diagram is considered to be a nonparametric estimator of the square root of two times the circular density. The mean integrated square error of the rose diagram and its optimal bin width is derived, and two new selectors are proposed: normal reference rule and biased cross-validation. The normal reference rule is a parametric method assuming an underlying density is a von Mises density. Biased cross-validation is a nonparametric method without the specification of an underlying density. It is shown that biased cross-validation converges to its optimizer. The numerical experiment is conducted under some simulation scenarios based on sine-skewed distributions to investigate how a choice of bin width affects the performance of the rose diagram under finite samples. Its result shows that biased cross-validation or normal reference rule outperforms some previous selectors and biased cross-validation has the best performance.

E0609: **New family of a toroidal distribution whose marginal and conditional distributions are skewed**

Presenter: **Tomoaki Imoto**, University of Shizuoka, Japan

Circular distribution is used for analyzing data whose observations are represented as points in the circumference of a unit circle, which is called circular data. In many cases, circular observations are made on two or more circular variables, like a combination of the directions of the wind at some points. For modelling and analyzing such data, distribution on the torus, called toroidal distribution, is a useful tool. The family of a sine-correlated distribution is a simple model, and its marginal distributions can be specified by any circular distribution. However, the circular correlation structure is too weak, and therefore, its applications will be restricted. A new method is proposed for constructing a toroidal distribution whose correlation structure is strong. The marginal and conditional distributions are skewed distributions, and the trigonometric moments are expressed in a simple form. From the expression, the role of parameters in the constructed distribution will be found. The illustrative example of fitting real data to the proposed distribution is also shown.

E0792: **A skew transition distribution modeling for higher-order circular Markov processes**
Presenter: **Hiroaki Ogata**, Tokyo Metropolitan University, Japan
Co-authors: Takayuki Shiohama

The purpose is to propose an extension of the higher-order Markov processes on the circle where an underlying binding density has a skewing structure. The structures for circular autocorrelation functions (CACF), circular partial autocorrelation functions (CPACF), and the spectral density function of the process are investigated. The maximum likelihood estimation for model parameters is considered, and its finite sample performances are investigated by numerical simulations. As a real data analysis, time series of wind directions is used for practical purposes.

EO238 Room 705 SPATIAL STATISTICS **Chair: Debashis Mondal**

E0874: **A graph neural network approach for spatial statistical modeling**
Presenter: **Snigdhansu Chatterjee**, University Of Minnesota, United States
Co-authors: Somya Sharma

Artificial neural network-based approaches for modelling have been singularly successful for several applications, especially with image data and other datasets that have some similarities with spatial data. An algorithm that uses a graph neural network is presented for modelling spatial data. A Bayesian framework is imposed on the graph neural network for statistical inference and uncertainty quantification.

E0880: **High dimensional logistic regression under network dependence**
Presenter: **Somabha Mukherjee**, National University of Singapore, Singapore
Co-authors: Ziang Niu, Bhaswar Bhattacharya, George Michailidis, Sagnik Halder

The classical formulation of logistic regression relies on the independent sampling assumption, which is often violated when the outcomes interact through an underlying network structure, such as over a temporal/spatial domain or on a social network. This necessitates the development of models that can simultaneously handle both the network peer effect and the effect of high-dimensional covariates. A framework for incorporating is described such dependencies in a high-dimensional logistic regression model by introducing a quadratic interaction term designed to capture the pairwise interactions from the underlying network. The resulting model can also be viewed as an Ising model, where the node-dependent external fields linearly encode the high-dimensional covariates. A penalized maximum pseudo-likelihood method is used for estimating the network peer effect and the effect of the covariates (the regression coefficients), which conveniently avoids the computational intractability of the maximum likelihood approach. The results imply that even under network dependence, it is possible to consistently estimate the model parameters at the same rate as in classical logistic regression when the true parameter is sparse, and the underlying network is not too dense. The rates of consistency of the proposed estimator are also presented for various natural graphs ensembles, such as bounded degree graphs, sparse Erdos-Renyi random graphs, and stochastic block models.

E0903: **Spatial confounding in spatial linear mixed models**
Presenter: **Kori Khan**, Iowa State University, United States

In the last two decades, considerable research has been devoted to a phenomenon known as spatial confounding. Spatial confounding is thought to occur when there is collinearity between a covariate and the random effect in a spatial regression model. This collinearity is often considered highly problematic when the inferential goal is estimating regression coefficients and various methodologies have been proposed to "alleviate" it. Recently, it has become apparent that many of these methodologies are flawed, yet the field continues to expand. The purpose is to synthesise work in the field of spatial confounding. It is proposed that at least two distinct phenomena are currently conflated with the term spatial confounding. These are referred to as the analysis model and the data generation types of spatial confounding. In the context of spatial linear mixed models, it is shown that these two issues can lead to contradicting conclusions about whether spatial confounding exists and whether methods to alleviate it will improve inference. The results also illustrate that, in many cases, traditional spatial models help improve the inference of regression coefficients. Drawing on the insights gained, a path forward is offered for research in spatial confounding.

E0909: **Additive dynamic models for correcting numerical model outputs**
Presenter: **Xiaohui Chang**, Oregon State University, United States

Numerical air quality models are pivotal for predicting and assessing air pollution, but numerical model outputs may be systematically biased. An additive dynamic model is proposed to correct large-scale raw model outputs using data from other sources, including readings collected at ground monitoring networks and weather outputs from different numerical models. An additive partially linear model specification is employed for the nonlinear relationships between air pollutants and covariates. In addition, a multi-resolution basis function approximation is proposed to capture the different small-scale variations of biases. A discretized stochastic integrodifferential equation is constructed to characterize the dynamic evolution of the random coefficients at each spatial resolution. An expectation-maximization algorithm is developed for parameter estimation, and a multi-resolution ensemble-based scheme is embedded to accelerate the computation. The proposed approach is used to correct the biased raw outputs of PM2.5 from the Community Multiscale Air Quality (CMAQ) system for China's Beijing-Tianjin-Hebei region. The method improves the root mean squared error and continuous rank probability score by 43.7% and 34.76%, respectively. Compared to other statistical methods under different metrics, this model has correction accuracy and computational efficiency advantages.

EO240 Room 708 ADVANCES IN FUNCTIONAL DATA ANALYSIS AND THEIR APPLICATIONS **Chair: Hidetoshi Matsui**

E0497: **Spatiotemporal factor models for functional data with application to population map forecast**
Presenter: **Tomoya Wakayama**, The university of Tokyo, Japan
Co-authors: Shonosuke Sugasawa

With the proliferation of mobile devices, an increasing amount of population data is being collected, and there is growing demand to use large-scale, multidimensional data in real-world situations. Functional data analysis (FDA) was introduced into the problem of predicting the hourly population of different districts of Tokyo. FDA is a methodology that treats and analyzes longitudinal data as curves, which reduces the number of parameters and makes it easier to handle high-dimensional data. Specifically, by assuming a Gaussian process, the large covariance matrix parameters of the multivariate normal distribution were avoided. In addition, the data were time and spatially-dependent between districts. To capture these characteristics, a Bayesian factor model was introduced, which modelled the time series of a small number of common factors and expressed the spatial structure in terms of factor loading matrices. Furthermore, the factor-loading matrices were made identifiable and sparse to ensure the interpretability of the model. A method for selecting factors using the Bayesian shrinkage method is also proposed. The forecast accuracy and interpretability of the proposed approach through numerical experiments and data analysis were studied. It was found that the flexibility of the proposed method could be extended to reflect further time series feature contributing ted to the accuracy.

E0815: **On smoothing for spatial functional data**
Presenter: **Yoshikazu Terada**, Osaka University; RIKEN, Japan
Co-authors: Hidetoshi Matsui

With the recent advances in measurement technology, it has become easier to acquire spatiotemporal data, and the importance of analyzing spatiotemporal data is increasing in various fields. The smoothing (or interpolation) problem in spatial functional data is considered. Firstly, a unified view of the existing basis expansion approaches is provided. Then, a new, simple smoothing procedure for spatial functional data is proposed. The proposed method adopts spatial regularization for the basis coefficients, which induces both spatial and temporal smoothness. The proposed method's performance is compared with existing methods through numerical experiments.

E0843: **A functional generalized additive model-based scan statistic for disease cluster detection**
Presenter: **Michio Yamamoto**, Osaka University / RIKEN AIP, Japan
Co-authors: Tatsuhiko Anzai, Kunihiko Takahashi

Detecting spatial clusters of diseases is crucial for understanding disease patterns and developing effective prevention and treatment strategies. Spatial scan statistics are powerful methods for detecting spatial clusters with a variable scanning window size. A new spatial scan statistic is proposed that considers the spatial correlation of an outcome variable and handles multiple functional covariates that indicate past information over time. Specifically, the proposed method flexibly models these factors using the framework of functional generalized additive models. The performance of the proposed approach will be examined through a simulation study and real data analysis.

E0849: **Functional mixture cure model and its application**
Presenter: **Yuko Araki**, Tohoku University, Japan

A functional mixture cure model with functional covariates measured at the baseline period is considered. Much biostatistical research focuses on the time to event and its association with risk factors. The mixture cure model is based on the idea that the entire population can be divided into cure and un-cure groups. The cure group means that they do not develop the symptom of interest in the future. The mixture cure model has been used in the evaluation of the effects of therapeutic agents on cancer recurrence or a study of patients who did not die after recovering from Covid 19. However, the existing methods do not treat the situation that some covariates of interest during the baseline period are a function of time or space. A functional mixture cure model is proposed, which can deal with such data to reveal the association between risk factors and mortality with some complex changes in clinical measurements.

EO166 Room 709 HIGH-DIMENSIONAL/SPATIAL TIME SERIES ANALYSIS: THEORY AND APPLICATIONS Chair: Yasumasa Matsuda

E0689: **Estimation of single index models in moderately high dimension**
Presenter: **Kazuma Sawaya**, The University of Tokyo, Japan
Co-authors: Yoshimasa Uematsu, Masaaki Imaizumi

A new estimator is proposed for semiparametric single index models in the moderately high-dimensional regime, where the number of covariates p grows proportionally with the sample size n. Asymptotic unbiasedness and asymptotic normality of the estimator without the sparsity condition of the true parameter are also guaranteed. The estimation is based on the deconvolution method and the generalized approximate message-passing algorithm. Numerical simulations show the validity of our theory.

E0692: **Estimation of large covariance matrices with mixed factor structure**
Presenter: **Runyu Dai**, Tohoku University, Japan
Co-authors: Yoshimasa Uematsu, Yasumasa Matsuda

The Principal Orthogonal complEment Thresholding (POET) framework is extended to estimate large covariance matrices with a "mixed" structure of observable and unobservable strong/weak factors, and this method is called the extended POET (ePOET). Especially, the weak factor structure allows the existence of much slowly divergent eigenvalues of the covariance matrix frequently observed in real data. Under some mild conditions, the uniform consistency of the proposed estimator is derived for the cases with or without observable factors. Furthermore, several simulation studies show that the ePOET achieves good finite-sample performance regardless of data with strong, weak, or mixed factors structure. Finally, empirical studies are conducted to present the practical usefulness of the ePOET.

E0818: **Asset pricing with co-search interaction**
Presenter: **Stanley Iat-Meng Ko**, Tohoku University, Japan

The effect of internet co-search activities of listed stocks on their returns in the stock market is studied. The internet traffic is explored on the US Securities and Exchange Commission (SEC) Electronic Data Gathering, Analysis, and Retrieval (EDGAR) website, which holds all public US companies' information with hundreds of thousands of document views per day by users. Co-searched firms are identified, i.e. one firm is searched subsequently after another, and such information is incorporated into the conventional asset pricing model. First, the micro-level behavioural information of individual stocks is introduced to the empirical asset pricing literature, whereas traditional asset pricing studies focus on aggregated portfolios. Second, with the identification of co-search peers, the co-search network is defined and constructed in the universe of trading stocks. The virtual spatial stock return dependence across the network is identified through the co-search network lens. Third, the traditional liner asset pricing models are extended using the Spatial Arbitrage Pricing Theory (S-APT).

E0694: **Deep learning for multivariate volatility forecasting in high-dimensional financial time series**
Presenter: **Yasumasa Matsuda**, Tohoku University, Japan
Co-authors: Rei Iwafuchi

Volatility modeling is considered for high-dimensional financial time series by a long short-term memory (LSTM) neural network. We apply a deep LSTM neural network to describe multivariate volatility dynamic behaviours in financial time series. We discuss the empirical features of the LSTM modeling for the SP500 return series in comparison with those of popular existing models in terms of volatility forecasting performances.

EC321 Room 04 MULTI-STATE AND COX MODELS Chair: Russell Shinohara

E0233: **A Bayesian approach for chronic hepatitis C prevalence estimation to improve the accuracy of economic evaluation**
Presenter: **William WL Wong**, University of Waterloo, Canada
Co-authors: Zeny Feng

The majority of those infected with chronic hepatitis C (CHC) have a clinical silent disease. The asymptomatic nature means the disease often remains undiagnosed, leaving its prevalence highly uncertain. This generates significant uncertainty for the associated economic evaluations. The purpose is to establish a mathematical framework for estimating CHC prevalence and undiagnosed proportion. A state-transition model describing infection, disease progression and treatment response was mathematically formulated and developed. Model parameters were obtained from the published literature. The historical prevalence of CHC is estimated through a calibration process based on a Bayesian MCMC algorithm. The algorithm constructed posterior distributions of the historical prevalence of CHC by comparing the model-generated predictions of the annual numbers of health events related to CHC against the observed calibration targets. The prevalence of CHC in Ontario, Canada, in 2018 was estimated to be 0.89%, and the percentage of undiagnosed among the total infected was 33.6%. The results are in line with a recently conducted seroprevalence survey. Prevalence estimates impact economic evaluation results on interventions concerning CHC screening and treatment. Considering the rapid development of treatments for CHC, updated prevalence estimates will become necessary. A platform is provided for estimating this information in a robust and efficient way.

E0289: **Risk factors and transitional probability of clinical events in Korean CKD patients using the multi-state model**
Presenter: **Jinheum Kim**, University of Suwon, Korea, South

Since the KNOW-CKD study is a longitudinal study, CKD patients are not only exposed to fatal events such as death during the study period but also to non-fatal events such as cardiovascular disease (CVD) and end-stage kidney disease (ESKD). Such non-fatal events are called intermediate events. By ignoring intermediate events, estimating the survival probability of CKD patients only with information on whether they survive or are censored may lead to misleading results. In situations where intermediate events can be experienced in a patient's disease progression, do not consider a two-state model consisting of only the initial state and death state as an analytical model but use so-called multi-state models created by adding intermediate states to the two-state model. It is better to consider the most used multi-state model is the so-called illness-death model, which consists of a healthy state corresponding to an initial state, an illness state corresponding to an intermediate event, and a death state corresponding to an absorbing state. The advantage of this illness-death model over the two-state model is that it can compare the effects of illness on patient mortality. A four-state model that extends the illness-death model is proposed to compare the effects of CVD or ESKD experience on patient death. Through this multi-state model, the disease progression of CKD patients can be explained more adequately.

E1196: **Tensor association test in Cox regression model**
Presenter: **Chin-Chun Chen**, National Cheng Kung University, Tainan, Taiwan
Co-authors: Pei-Fang Su

In survival analysis, representing multi-omics data as a tensor structure diversifies the types of covariates in the Cox regression model. Due to the complexity and high dimensionality of tensor covariates, a tensor rank decomposition technique is utilized for dimensionality reduction. A block relaxation algorithm, based on partial likelihood, is developed to iteratively estimate the effects of both clinical and tensor covariates. The innovation is a tensor association test, which facilitates statistical inference between survival outcomes and the two types of covariates. The proposed test enables the identification of genes within different platforms that demonstrate significant relationships with the survival outcomes. A comprehensive simulation study and real-case study on the Colorectal Adenocarcinoma TCGA PanCancer data are applied to demonstrate the method.

E1199: **Synthesizing auxiliary subgroup restricted mean survival information to obtain efficient estimation for Cox model**
Presenter: **Jo-Ying Hung**, National Cheng Kung University, Taiwan
Co-authors: Pei-Fang Su

Restricted mean survival time (RMST) has gained attention in the medical research field because it is model-free and easy to interpret. Due to the increasing accessibility of data sources, it is of interest to make use of published external information, such as RMST, to increase the efficiency of estimators in individual-level studies. In this research, a double empirical likelihood method is proposed that incorporates auxiliary RMST into the estimation of the Cox proportional hazard model. It is proved that the proposed estimator asymptotically follows a multivariate normal distribution and is asymptotically more efficient than the classical partial likelihood (PL) estimator. Simulation studies show that the proposed method yields more efficient estimators compared to the PL estimators. A diabetes dataset is used to demonstrate the proposed method.

EC261 Room 201 METHODOLOGICAL STATISTICS AND ECONOMETRICS Chair: Tadao Hoshino

E0276: **A score-driven filter for time-varying regression models with endogenous regressors**
Presenter: **Noah Stegehuis**, Vrije Universiteit Amsterdam, Netherlands
Co-authors: Francisco Blasques

A score-driven model is proposed for filtering time-varying causal parameters using instrumental variables. The causal parameter is updated at each time point by the score of the likelihood function. In the presence of suitable instruments, it is shown that it can uncover dynamic causal relations between variables, even in the presence of regressor endogeneity which may arise due to simultaneity, omitted variables, or measurement errors. Due to the observation-driven nature of score models, the method is simple and practical to implement. The asymptotic properties of the maximum likelihood estimator are established and show that the instrumental variable score-driven filter converges to the unique unknown causal path of the true parameter, whereas the existing score-driven procedure does not. Further, the finite sample properties of the filtered causal parameter in a comprehensive Monte Carlo exercise are analysed. Finally, the empirical relevance of this method in an application is revealed to aggregate consumption in macroeconomic data.

E0362: **Model selection in panel data model with large number of fixed effects**
Presenter: **Zhaoyuan Li**, The Chinese University of Hong Kong, Shenzhen, China

Due to the bad control problem, fixed effects are used to replace control variables in panel regression models. However, a large number of fixed effects will lead to a false positive problem. Suppose that an insignificant result is obtained from a panel model, but a significant result may come up by adding more fixed effects. One question is how to decide which one is correct. And then, what can we do if it is a false positive? A new model selection approach is proposed to answer these two questions.

E1258: **Doubly robust estimation of average partial effects**
Presenter: **Harvey Klyne**, University of Cambridge, United Kingdom
Co-authors: Rajen D Shah

Single-parameter summaries of variable effects are desirable for ease of interpretation, but linear modelling assumptions commonly result in poor model fitting. A modern approach is to estimate the average partial effect—the average slope of the regression function with respect to the predictor of interest—using a doubly robust semiparametric procedure. Existing work has focused on specific forms of nuisance function estimators. The scope is extended to arbitrary plug-in nuisance function estimation, allowing for the use of modern machine learning methods. The procedure involves resmoothing a first-stage regression estimator to produce a differentiable version and modelling the joint distribution of the predictors through a location-scale model. It is proven that the proposals lead to a semiparametric efficient estimator under weak assumptions, and attractive numerical performance is demonstrated even under misspecification.

E0516: **Assessing heterogeneity in treatment effects**
Presenter: **Tetsuya Kaji**, University of Chicago, United States
Co-authors: Jianfei Cao

Treatment heterogeneity is of significant concern in economics, but the lack of identification of the joint distribution of the treated and control outcomes hinders its assessment. For example, the effect of having insurance on the health of otherwise unhealthy individuals may need to be assessed, but it is often infeasible to ensure only the unhealthy ones, and thus the causal effects for those are not identified. , here may be interested in the shares of winners and losers from a minimum wage increase, but the shares are not identified without the joint distribution. It is shown that these quantities are partially identified and derive tight bounds that complement quantile treatment effects.

EC329 **Room 203** **TIME SERIES I** **Chair: Kaiji Motegi**

E0328: **Threshold models for high-dimensional nonlinear time series**
Presenter: **Chi Tim Ng**, Hang Seng University of Hong Kong, Hong Kong
Co-authors: Yan Wu, Yuanbo Li

The threshold autoregressive (TAR) model is an important class of nonlinear time-series models and has attracted great attention in the literature. To apply the TAR model in high-dimensional settings, a threshold network autoregressive (TNAR) model is proposed to overcome the over-parameterisation difficulty by exploiting the network relations' available information. The sufficient conditions for the strict stationarity and the ergodicity of the TNAR model are established. A computationally efficient method is developed to estimate the multiple thresholds and the parameters. Grouped TNAR model is also proposed to reduce the number of parameters further. The asymptotic behaviour of the adaptive fused LASSO is explored, and the estimation consistency of both numbers of groups and group membership structure is established. The applicability of the proposed models is illustrated with the U.S. stock data.

E1158: **Measuring accordance movement between time series by Kendalls tau**
Presenter: **Ying Zhang**, Acadia University, Canada

Choosing accordance/similarity measures between time series is critical in time series data science applications. There are different types of similarity measures, from traditional statistical correlation coefficients to measurements recently developed by computer scientists. Due to the lack of distribution properties, many of such measures have no statistical inference power. Motivated by an investigation of evaluating the performance of a gas multi-sensor device for monitoring urban pollution, we focus on measuring the accordance movement and its inference between a signal process to a reference process based on the method of Kendall Tau. Kendall Tau-based coefficients in measuring time series accordance movement are defined; it is shown how to make inferences by constructing conference intervals, and finally, the method with the pollution data from the gas multi-sensor device described above is demonstrated.

E0369: **Burn-in selection in simulating time series**
Presenter: **Chun Yip Yau**, Chinese University of Hong Kong, Hong Kong

Many time series models are defined in a recursive manner, prohibiting exact simulations. In practice, one appeal to simulating a long time series and discarding a large number of initial simulated observations, known as the burn-in. For autoregressive models where the dependence decays exponentially fast, the choice of the burn-in is not critical. However, it is unclear how to select the burn-in number for long-memory time series where the dependence on the remote past is strong. By combining several samplers with randomized burn-in numbers, a method for exactly simulating the expectation of a statistic computed from a time series is developed. Moreover, with some suitably chosen statistics, the exact simulation method can be applied to quantify the effect of burn-in numbers on the simulated sample. Simulation studies are conducted to provide some practical guidance for burn-in selections.

E1322: **Forecasting and change point test for nonlinear heteroscedastic time series based on support vector regression**
Presenter: **Meihui Guo**, National Sun Yat-sen University, Taiwan

SVR-ARMA-GARCH models provide flexible model fitting and good predictive powers for nonlinear heteroscedastic time series datasets. The change point detection problem in the SVR-ARMA-GARCH model using the residual-based CUSUM test is explored. For this task, an alternating recursive estimation (ARE) method is proposed to improve the estimation accuracy of residuals. Moreover, using a new testing method with a time-varying control limit significantly improves the detection power of the CUSUM test is suggested. The numerical analysis exhibits the merits of the proposed methods in SVR-ARMA-GARCH models. A real data example is also conducted using BDI data for illustration, which also confirms the validity of the methods.

Tuesday 01.08.2023 16:30 - 18:35 Parallel Session E – EcoSta2023

EV257 Room 701 FINANCIAL ECONOMETRICS (VIRTUAL) Chair: Toshiaki Watanabe

E1035: **Uncertainty and volatility: A Markov-switching GARCH-MIDAS approach**
Presenter: **Yao Rao**, The University of Liverpool, United Kingdom

A flexible specification is proposed that encompasses virtually all GARCH-MIDAS models proposed in the literature. The specification accounts for asymmetry and regime switches in the volatility dynamics. In the empirical application, the relationship between S&P 500 returns volatility and macroeconomic uncertainty is examined. It is shown that the model provides more accurate volatility forecasts than the nested GARCH-MIDAS models at forecast horizons of 1 day, two weeks, one month, two months and three months. Furthermore, the findings suggest that while high-frequency uncertainty indices are suitable for short-horizon forecasts, low-frequency uncertainty proxies provide better forecasts at longer horizons.

E1250: **Real-time indicator of financial fragmentation in the euro area**
Presenter: **Roland Bouillot**, Catholic University of Louvain, Belgium
Co-authors: Bertrand Candelon, Iason Kynigakis

Financial fragmentation in the Euro Area has become a hot topic since the European Central Bank must decide whether to maintain its fund's rates high and long enough to tame inflation or loosen its monetary policy to mitigate the risk of another European sovereign debt crisis. By using mixed-frequency high-dimensional data, the inter-country shock transmission is investigated through regularization techniques and vector autoregressive models to create a new real-time indicator of financial fragmentation. It is expected to find evidence of financial fragmentation resurgence during COVID-19 due to the pressure on public finances through the increase in sovereign debt levels and the widening of public deficits. Most importantly, the financial fragmentation risk is expected to persist and even intensify due to the ECB rate hikes amid the recent inflation burst period. Hence, monitoring this new real-time indicator tightly is an effective way to keep track of the impending financial fragmentation risk and contributes to the surveillance of the Euro Area's financial stability.

E1052: **Diversifying risk parity portfolios with high-frequency principal components**
Presenter: **Laura Garcia-Jorcano**, Universidad de Castilla-La Mancha, Spain
Co-authors: Massimiliano Caporin, Juan-Angel Jimenez-Martin

The diversified risk parity (DRP) strategy for multi-asset allocation is used for generating diversified equity portfolios. Creating uncorrelated risk sources by means of high-frequency principal components analysis (HF-PCA), maximum diversification portfolios are obtained when equally budgeting risk to each of the uncorrelated risk sources. Especially the risk factors are forecasted, and the role of firms/industries is traced as potential sources of financial risk in different periods of time. The empirical analysis carried out using one-minute returns of stocks included in the S&P 100 index from 2003 to 2022 belonging to ten industry groups shows that compared to classical risk-based allocation schemes, the DRP strategy provides the most convincing risk-adjusted performance and the most diversified portfolio among the investigated alternatives according to several concentration indices and risk decomposition characteristics. HF-PCA allows the DRP strategy to constantly adapt to changes in risk structure and maintain a balanced exposure to the then prevailing uncorrelated risk sources. This tool can help a portfolio manager to understand and choose those risk sources that have earned risk focusing on those risk factors.

E1252: **Institutional stock-bond portfolios rebalancing and financial stability**
Presenter: **Jean-Baptiste Hasse**, Aix-Marseille University, France
Co-authors: Souhila Siagh, Christelle Lecourt

Rebalancing options are examined for long-term institutional investors. Specifically, risk-adjusted performances of different stock-bond portfolios between buy-and-hold, periodic and threshold rebalancing are estimated. Using the Norwegian Sovereign Wealth Fund (SWF) as a case study and an econometric approach based on a bootstrap test of Sharpe ratios difference, it is shown that the optimal rebalancing differs across economic and financial cycles. Furthermore, it is found that the optimal rebalancing is quarterly rebalancing except during recessions and crises when buy-and-hold is best, thus calling into question the hypothesis of the countercyclical behaviour of SWFs. The results are robust to alternative performance measures, asset allocations, investment horizons, non-normal returns, transaction costs and time sampling. Last, the findings promote the consideration for macro-prudential rules to improve the Santiago Principles and a specific monitoring framework targeted at the SWFs.

E0200: **Understanding growth-at-risk: A Markov-switching approach**
Presenter: **Francesca Loria**, Federal Reserve Board, United States

It is shown that a Markov-switching model with endogenous transition probabilities can replicate a common finding in the Growth-at-Risk literature: that the (conditional) mean, and volatility of future growth are negatively correlated. The model also provides an intuitive interpretation of macroeconomic risk: (endogenous) regime uncertainty generates tail risk. The higher the regime uncertainty, the starker the differences in the growth outlook between a normal and a bad state of the economy. The model is a new tool to assess the risk of tail events, such as recessions, and to evaluate the likelihood of point forecasts. Real-time measures of the United States' financial conditions and economic activity are also proposed. These measures are used to construct conditional quantiles and predictive distributions of average GDP growth over the next 12 months. It is shown that periods of high macroeconomic 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.

EO138 Room 02 LEARNING METHODS FOR LATENT STRUCTURES Chair: Long Nguyen

E0350: **Particle variational Bayes**
Presenter: **Minh-Ngoc Tran**, University of Sydney, Australia
Co-authors: Paco Tseng, Robert Kohn

Variational Bayes (VB) is widely recognised as a highly efficient and scalable technique for Bayesian inference. However, classical VB often imposes restrictions on the space of variational distributions, typically restricting it to a specific set of parametric distributions or factorised distributions. Ways to relax these restrictions by traversing a set of particles are explored to approximate the target distribution. The theoretical basis of this Particle VB method can be established using the Optimal Transport theory, which allows us to make the space of probability measures into a differential manifold. The focus is particularly on the novel Particle Mean Field Variational Bayes (PMFVB) approach, which extends the classical MFVB method without requiring conjugate priors or analytical calculations. The theoretical basis of the new method by leveraging the connection between Wasserstein gradient flows and Langevin diffusion dynamics is established. The effectiveness of this approach is demonstrated using Bayesian logistic regression, stochastic volatility, and deep neural networks.

E0489: **Harnessing geometric signatures in causal representation learning**
Presenter: **Yixin Wang**, University of Michigan, United States

Causal representation learning aims to extract high-level latent factors from low-level sensory data. Existing methods often identify these latent factors by assuming they are statistically independent. However, correlations between latent factors are prevalent across applications. It is explored how geometric signatures of latent causal factors can facilitate causal representation learning without any assumptions about their distributions or dependency structure. The key observation is that the absence of causal connections between latent causal factors often carries geometric signatures of the latent factors' support (i.e. what values each latent can possibly take). Leveraging this fact, latent causal factors are identified for permutation and scaling with data from perfect do interventions. Moreover, block affine identification with data can be achieved from imperfect interventions. These results highlight the unique power of geometric signatures in causal representation learning.

E0695: **Convergence rates for softmax gating Gaussian mixtures of experts**
Presenter: **Nhat Pham Minh Ho**, University of Texas, Austin, United States

Gaussian mixtures of experts with softmax gating functions have been used successfully in numerous applications, including computer vision, speech recognition, system identification, and recently large language models (e.g., Transformer). Despite their popularity, a comprehensive understanding of the behaviours of parameter estimation in these models has remained elusive. The maximum likelihood estimation (MLE) convergence rates are established for these models. The results indicate that the rates of MLE can be (very) slow due to an intrinsic interaction between the expert functions and the softmax gating functions. Finally, based on insights from the theory, a new variant of softmax gating functions that yields much faster convergence rates of parameter estimation in Gaussian mixtures of experts is proposed.

E0739: **Gibbs sampling for mixtures in order of appearance: The ordered allocation sampler**
Presenter: **Pierpaolo De Blasi**, University of Torino and Collegio Carlo Alberto, Italy
Co-authors: Maria Fernanda Gil-Leyva Villa

Gibbs sampling methods are standard tools to perform posterior inference for mixture models. These have been broadly classified into two categories: marginal and conditional methods. While conditional samplers are more widely applicable than marginal ones, they may suffer from slow mixing in infinite mixtures, where some form of truncation, either deterministic or random, is required. In mixtures with a random number of components, the exploration of parameter spaces of different dimensions can also be challenging. These issues are tackled by expressing the mixture components in the random order of appearance in an exchangeable sequence directed by the mixing distribution. A sampler that is straightforward is derived from implementing mixing distributions with tractable size-biased ordered weights, and that can be readily adapted to mixture models for which marginal samplers are not available. In infinite mixtures, no form of truncation is necessary. As for finite mixtures with random dimensions, a simple updating of the number of components is obtained by a blocking argument, thus, easing challenges found in transdimensional moves via Metropolis-Hastings steps. Additionally, sampling occurs in the space of ordered partitions with blocks labelled in the least element order, which endows the sampler with good mixing properties. The performance of the proposed algorithm is evaluated in a simulation study. Supplementary materials for this article are available online.

E0891: **Difficulties and nonstandard minimax rates in nonparametric latent variable models and representation learning**
Presenter: **Bryon Aragam**, University of Chicago, United States

One of the key paradigm shifts in statistical machine learning over the past decade has been the transition from handcrafted features to automated, data-driven representation learning, typically via deep neural networks. As these methods are being used in high-stakes settings such as medicine, health care, law, and finance, where accountability and transparency are not just desirable but often legally required, it has become necessary to place representation learning on a rigorous scientific footing. The statistical foundations of nonparametric latent variable models are revisited and discussed how even basic statistical properties such as identifiability and consistency are surprisingly subtle. New results are also discussed, characterizing the optimal sample complexity for learning simple nonparametric mixtures, which turn out to have a nonstandard super-polynomial bound. With time permitting, applications will end to deep generative models widely used in practice.

EO148 Room 03 INNOVATIVE DESIGN AND ANALYSIS METHODS FOR OPTIMIZING HEALTHCARE Chair: Nina Deliu

E0741: **Contextual bandits for online decision-making in mobile health studies with count proximal outcomes**
Presenter: **Xueqing Liu**, Duke-NUS Medical School, Singapore
Co-authors: Nina Deliu, Bell Lauren, Bibhas Chakraborty

Mobile health (mHealth) technologies aim to improve distal outcomes, such as clinical conditions, by optimizing proximal outcomes through just-in-time adaptive interventions. Contextual bandits provide a suitable framework for customizing such interventions, with an agent continuously observing contexts and choosing from available interventions to maximize cumulative proximal outcomes. However, unique challenges, such as modelling count outcomes within bandit frameworks, have hindered the widespread application of contextual bandits to mHealth studies. This challenge is addressed by leveraging count data models into online decision-making approaches. Specifically, four common offline count data models (Poisson, negative binomial, zero-inflated Poisson, and zero-inflated negative binomial regressions) are combined with Thompson sampling, a popular contextual bandit algorithm. The proposed algorithms are motivated by and evaluated on a real dataset from the Drink Less trial, where they are used to improve user engagement by increasing the number of screen views compared to existing methods. The proposed methods are further evaluated on simulated data, achieving improvement in maximizing cumulative proximal outcomes over existing algorithms. Theoretical results on convergence rate and regret bound are also derived.

E0723: **Missing data: A key challenge for digital outcomes in clinical trials**
Presenter: **Mia Sato Tackney**, University of Cambridge, United Kingdom

Clinical trials increasingly use digital devices to measure the impact of an intervention on health outcomes. For example, accelerometers in physical activity trials can measure step count in very fine intervals of time, typically in 5-second epochs. The data is usually aggregated to provide the primary analysis's daily or weekly step counts. Missing data is common as participants may not wear the device per protocol, or there may be device failure. Approaches to handling missing data in the literature have largely defined missingness on the day level using a threshold on wear time, which leads to loss of information on the time of day when data is missing. An approach is presented to identifying and classifying missingness at the finer epoch level. Missingness can then be handled using a non-parametric approach to Multiple Imputation (MI), where missing periods during the day are replaced by donor data from the same person where possible or data from a different person who is matched on demographic variables. An application of this approach is illustrated in the 2017 PACE-UP Trial. Finally, the key statistical opportunities and challenges to adopting digital endpoints in clinical trials are discussed.

E0766: **Bayes optimal decision-making in adaptive enrichment clinical trials**
Presenter: **Thomas Burnett**, University of Bath, United Kingdom

Adaptive Enrichment designs allow for the selection of pre-defined patient sub-groups at a pre-planned interim analysis. Given the possible pathways through the trial are pre-defined, hypothesis tests may be defined to ensure strong control of the FamilyWise Error Rate for every possible interim decision. This structure allows complete freedom for how these interim decisions are made. A Bayesian decision framework is used to find optimal decisions in the interim analysis. Evaluating the overall performance of these Bayes optimal Adaptive Enrichment trials is computationally intensive. An algorithm for computing the form of the optimal decision rule shall be discussed. This allows a comparison of the adaptive designs with fixed sampling alternatives, showing the possible benefits of the adaptive methods. Further to optimising the interim analysis, the same optimisation principle may be applied to the whole trial. By optimising across the two-stage adaptive design, fully optimal trials and gain further efficiency in the adaptive design are found. The fixed sampling designs and the previous Enrichment design are special cases in this optimisation, so as a direct result, the most efficient trial design is chosen.

E0754: **SMART-EXAM: Incorporating participants welfare into sequential multiple assignment randomized trials**
Presenter: **Xinru Wang**, Duke-NUS Medical School, Singapore
Co-authors: Nina Deliu, Yusuke Narita, Bibhas Chakraborty

Dynamic treatment regimes (DTRs) are sequences of decision rules that recommend treatments based on patients time-varying clinical conditions. The sequential multiple assignment randomized trial (SMART) is an experimental design that can provide high-quality evidence for constructing optimal DTRs. Despite its relative simplicity of implementation and desirable performance in comparing embedded DTRs, the SMART with balanced randomization is faced with inevitable ethical issues, including assigning many participants to the observed inferior treatment or the treatment they dislike, which might slow down the recruitment procedure and lead to higher attrition rates. In this context, a SMART is proposed under the Experiment-as-Market framework (SMART-EXAM), a novel SMART design that can potentially improve patient welfare by incorporating participant preferences and predicted treatment effects into the randomization procedure. The procedure of conducting a SMART-EXAM is described, and its theoretical and empirical statistical properties compared with other competing SMART designs are evaluated. The results indicate that the SMART-EXAM design can improve the welfare of participants enrolled in the trial while achieving a comparable ability to construct an optimal DTR. Finally, the practical potential of the SMART-EXAM design is illustrated using data from a SMART for children with attention deficit/hyperactivity disorder (ADHD).

E0858: **Unravelling timing and determinants of childhood vaccination: The Italian case**
Presenter: **Chiara Chiavenna**, Bocconi University, Italy
Co-authors: Filippo Trentini, Catia Rossana Borriello, Michele Ercolanoni, Giuseppe Preziosi, Pierluca Colacioppo, Danilo Cereda, A. Melegaro

Vaccination coverage (VC) is an important proxy of a population's level of protection against a specific pathogen. When VC is below a given threshold, public health officers implement policies encouraging uptake. The population subgroups that an optimal intervention should target are identified. Routine childhood vaccination records of 860,677 Lombardy residents born between 2006 and 2019 were extracted. Sequence analysis was used to summarize their vaccination behaviour: vaccination timings are expressed as life trajectories, with individual states indicating the number of doses received at each time unit. A distance matrix quantifying dissimilarity between sequences was used to cluster individuals; finally, a multivariate regression model was trained to predict cluster classification based on parental age, education, employment, nationality, and COVID-19 vaccination. For the hexavalent vaccine, four clusters of individuals were identified, labelled as vaccinated on time (90.19%), delayed vaccination (6.69%), interrupted the cycle (1.72%), and did not vaccinate (1.39%). For measles, higher proportions of children interrupted the cycle (4.15%) or did not vaccinate (2.35%). A higher probability of delaying or refusing vaccination was estimated for children with mothers unvaccinated against COVID-19. Children with at least one foreign parent or unemployed fathers were also at higher risk of non-compliance.

EO234 Room 04 RECENT ADVANCEMENTS IN RELIABILITY AND QUANTILE MODELING Chair: Tony Sit

E0965: **On reliability analysis of one-shot device testing data with defects**
Presenter: **Man Ho Ling**, The Education University of Hong Kong, Hong Kong

The problem of defective devices in the manufacturing industry is studied. Defective devices can arise due to a range of reasons, such as errors made by workers, inadequate quality processes, insufficient training, or issues with reliability during the design stage. The focus is on one-shot device test data that includes defects which have occurred during a realistic manufacturing process. In this scenario, the question arises of whether a failed device is actually defective or has simply failed due to its lifetime being shorter than the inspection time. The maximum likelihood approach will be explored to estimate the mean time to failure based on a sample of one-shot devices with manufacturing defects. This will be done using the gamma and Weibull lifetime distributions. How masking can affect the estimation and analysis under different defective rates and masking proportions will also be examined.

E1125: **On interquantile smoothness of censored quantile regression with induced smoothing**
Presenter: **Tony Sit**, The Chinese University of Hong Kong, Hong Kong

Quantile regression has emerged as a useful and effective tool in modelling survival data, especially for cases where noises demonstrate heterogeneity. Despite recent advancements, non-smooth components involved in censored quantile regression estimators may often yield numerically unstable results, which, in turn, lead to potentially self-contradicting conclusions. An estimating equation-based approach is proposed to obtain consistent estimators of the regression coefficients of interest via the induced smoothing technique to circumvent the difficulty. The proposed estimator can be shown to be asymptotically equivalent to its original unsmoothed version, whose consistency and asymptotic normality can be readily established. Extensions to handle functional covariate data and recurrent event data are also discussed. To alleviate the heavy computational burden of bootstrap-based variance estimation, an efficient resampling procedure is also proposed that reduces the computational time considerably. The numerical studies demonstrate that the proposed estimator provides substantially smoother model parameter estimates across different quantile levels and can achieve better statistical efficiency than a plain estimator under various finite-sample settings. The proposed method is also illustrated via four survival datasets, including the HMO HIV data, the primary biliary cirrhosis (PBC) data, etc.

E1306: **The gender wage gap over the life cycle: Evidence from Japan**
Presenter: **Jauer Chen**, Senshu University, Japan

The gender wage gap is a persistent and pervasive issue that has received significant attention from economists, policymakers, and the general public. Despite efforts to close the gap, it remains a challenge to quantify and understand the sources of this disparity. The aim is to shed light on the gender wage gap over the life cycle, using data from Japan's Basic Survey on Wage Structure (BSWS). This data provides rich information on the wages of employees and offers a unique opportunity to examine the pattern of the gender wage gap across different age groups. Our findings show that the gender wage gap widens with age, particularly among those with higher education levels. This growing pattern is consistent with previous research, which found that the gender earnings gap widens over the life cycle and is greater among college graduates than others. Our results also indicate that the gap is larger in the middle of the wage distribution in each age group, except for those over 55. Valuable insights into the life cycle of the gender wage gap in Japan are provided, contributing to a deeper understanding of this issue. Primarily results are preliminarily presented, obtained from the quantile decomposition and the RIF decomposition in the context of unconditional quantile regressions.

E1307: **Censored interquantile regression model with time-dependent covariates**
Presenter: **Chi Wing Chu**, City University of Hong Kong, Hong Kong
Co-authors: Tony Sit

Conventionally, censored quantile regression stipulates a specific, pointwise conditional quantile of the survival time given covariates. Despite its model flexibility and straightforward interpretation, the pointwise formulation oftentimes yields rather unstable estimates across neighbouring quantile levels with large variances. In view of this phenomenon, a new class of quantile-based regression models with time-dependent covariates for censored data is proposed. The models proposed aim to capture the relationship between the failure time and the covariate processes of a target population that falls within a specific quantile bracket. The pooling of information within a homogeneous neighbourhood facilitates more efficient estimates hence a more consistent conclusion on the statistical significance of the variables concerned. This new formulation can also be regarded as a generalization of the accelerated failure time model for survival data in the sense that it relaxes the assumption of global homogeneity for the error at all quantile levels. Numerical studies demonstrate that the proposed estimator outperforms existing alternatives under various settings in terms of smaller empirical biases and standard deviations. A perturbation-based resampling method is also developed to reconcile the asymptotic distribution of the parameter estimates. Finally, consistency and weak convergence of the proposed estimator are established via empirical process theory.

E1323: **Statistical inference for lifetimes of one-shot devices with gamma frailty components**
Presenter: **Ping Shing Ben Chan**, The Chinese University of Hong Kong, Hong Kong

The inferential procedure for the lifetimes of one-shot devices with M components is studied. Assuming the component's lifetimes are exponentially distributed and the dependency among component lifetimes is governed by a frailty parameter which is assumed to be gamma distributed. The unconditional distribution of the component lifetimes will be derived. First, the time-censored samples are considered. An EM algorithm is proposed to find the maximum likelihood estimators of the unknown parameters. The algorithm is very easy to implement. The estimates of the variance-covariance of the estimators based on the missing information principle are also given. Then the results are extended to one-shot device data. The EM algorithm for the time-censored data has been modified to tackle the one-shot device data. A simulation study will then be conducted to examine the performance of the proposed algorithm. The estimators' biases and mean square errors under various settings will be presented. The coverage probabilities and average lengths of the confidence intervals constructed based on the asymptotic distribution of the estimators will also be given. Finally, an example is presented to illustrate the method of inference developed in the article.

EO094 **Room Virtual R01** **EXTREME VALUE ANALYSIS FOR COMPLEX DATA** **Chair: Gilles Stupfler**

E0824: **A de-randomization argument for estimating extreme value parameters of heavy tails**
Presenter: **Joseph Hachem**, Toulouse School of Economics, France
Co-authors: Gilles Stupfler, Abdelaati Daouia

In extreme value analysis, it has recently been shown that one can use a de-randomization trick to replace a random threshold in the estimator of interest with its deterministic counterpart to estimate several extreme risks simultaneously, but only in an i.i.d. context. The aim is to show how this method can be used to handle the estimation of several tail quantities (tail index, expected shortfall, distortion risk measures...) in general dependence/heteroskedasticity/heterogeneity settings under a weighted L^1 assumption on the gap between the average distribution of the data and the prevailing distribution. Particularly interesting examples of application include serially dependent but heteroskedastic frameworks.

E0873: **Extremes in high dimensions: Methods and scalable algorithms**
Presenter: **Marco Oesting**, University of Stuttgart, Germany
Co-authors: Johannes Lederer

The extreme-value theory has been explored in considerable detail for univariate and low-dimensional observations, but the field is still in an early stage regarding high-dimensional multivariate observations. The method focuses on Huesler-Reiss models and their domain of attraction, a popular class of models for multivariate extremes that exhibit similarities to multivariate Gaussian distributions. Novel estimators are devised for the parameters of this model based on score matching and equip these estimators with state-of-the-art theories for high-dimensional settings and exceptionally scalable algorithms. A simulation study is performed to demonstrate that the estimators can estimate a large number of parameters reliably and fast; for example, Huesler-Reiss models with thousands of parameters are shown that can be fitted within a couple of minutes on a standard laptop. A real data example on weather extremes illustrates their usefulness for applications.

E0892: **Hedge funds systemic risks: Which factors matter?**
Presenter: **Philippe Hubner**, HEC Liege, University of Liege, Belgium
Co-authors: Julien Hambuckers

An extreme value systemic risk model is extended to a regression context, where a set of covariates drives marginal tail indices of hedge funds and banks. As such, systemic risk contributions of hedge funds to the banking sector are allowed to be time-varying and fund-specific. Moreover, this formulation makes it possible to estimate these contributions by exploiting pooled time series of hedge funds returns, overcoming the short reporting periods in commercial databases. Then funds characteristics and market conditions indicate a high systemic threat, information of interest for regulators. Using a sample of around 5,000 funds, it is found that the systemic risk contribution of hedge funds increased after the 2008 crisis for all strategies, mainly driven by the increasing individual size of funds. Moreover, the investment strategy is found to be a determinant factor in explaining systemic risk intensity. In particular, Fixed Income hedge funds are found to be the highest systemic risk contributors. Finally, hedge funds' systemic risk is revealed to increase during periods of high uncertainty despite a contemporaneous decrease in extremal dependence on the banking sector.

E0948: **Inference for extremal regression with dependent heavy-tailed data**
Presenter: **Gilles Stupfler**, University of Angers, France
Co-authors: Abdelaati Daouia, Antoine Usseglio-Carleve

Nonparametric inference on tail conditional quantiles and their least squares analogues, expectiles, remains limited to i.i.d. data. A fully operational inferential theory is developed for extreme conditional quantiles and executiles in the challenging framework of strong mixing, conditional heavy-tailed data whose tail index may vary with covariate values. This requires a dedicated treatment to deal with data sparsity in the far tail of the response, in addition to handling difficulties inherent to mixing, smoothing, and sparsity associated with covariate localization. The pointwise asymptotic normality of the estimators is proved, and optimal convergence rates are obtained reminiscent of those found in the i.i.d. regression setting but which had not been established in the conditional extreme value literature. The assumptions hold in a wide range of models. Full bias and variance reduction procedures and simple but effective data-based rules for selecting tuning hyperparameters are proposed. The inference strategy is shown to perform well in finite samples and is showcased in applications to stock returns and tornado loss data.

E0990: **Marginal expected shortfall inference under multivariate regular variation**
Presenter: **Matteo Schiavone**, Bocconi University, Italy
Co-authors: Simone Padoan, Stefano Rizzelli

Marginal expected shortfall is unquestionably one of the most popular systemic risk measures. Studying its extreme behaviour is particularly relevant for risk protection against severe global financial market downturns. In this context, the results of statistical inference rely on the bivariate extreme values approach, disregarding the extremal dependence among a large number of financial institutions that make up the market. To take it into account, an inferential procedure is proposed based on the multivariate regular variation theory. An approximating formula is derived for the extreme marginal expected shortfall and obtained from it an estimator and its bias-corrected version. Then, their asymptotic normality is shown, which allows, in turn, the confidence intervals derivation. Simulations show that the new estimators greatly improve upon the performance of existing ones, and confidence intervals are very accurate. An application to financial returns shows the utility of the proposed inferential procedure. Statistical results are extended to a general β-mixing context that allows working with popular time series models with heavy-tailed innovations.

EO226 Room Virtual R02 ASYMPTOTIC STATISTICS FOR STOCHASTIC PROCESSES Chair: Masayuki Uchida

E1080: **Quasi-likelihood analysis and estimation for a degenerate diffusion process**
Presenter: **Nakahiro Yoshida**, University of Tokyo, Japan

The quasi-likelihood analysis (QLA) is an abstract framework for statistical inference for stochastic processes. Under easily verified conditions, this theory enables us to obtain a polynomial-type large deviation inequality for the quasi-likelihood random field and, consequently, asymptotic properties, including moments convergence, of the quasi-maximum likelihood estimator and the quasi-Bayesian estimator. This scheme has been used for various parametric models of nonlinear stochastic processes thanks to its universal design and easy handling. A simplified quasi-likelihood analysis is applied to the parametric estimation of a degenerate diffusion process.

E0872: **Scaling limit of Markov chain/process Monte Carlo methods**
Presenter: **Kengo Kamatani**, ISM, Japan

The scaling limit analysis of Markov Chain Monte Carlo methods has been a topic of intensive study in recent decades. The analysis determines the rate at which the Markov Chain converges to its limiting process, typically a Langevin diffusion process, and provides valuable guidelines for parameter tuning. Numerous researchers have generalized the original assumptions and expanded the results to more sophisticated methods. Recently, there has been growing interest in piecewise deterministic Markov processes for Monte Carlo integration methods, particularly the Bouncy Particle Sampler and the Zig-Zag Sampler. The method focuses on determining the scaling limits for both algorithms and provides a criterion for tuning the Bouncy Particle Sampler.

E0351: **Parametric estimation for discretely observed linear parabolic SPDEs in two space dimensions**
Presenter: **Masayuki Uchida**, Osaka University, Japan
Co-authors: Yozo Tonaki, Yusuke Kaino

Estimating unknown coefficient parameters of linear parabolic second-order stochastic partial differential equations (SPDEs) in two space dimensions driven by Q-Wiener processes based on high-frequency data in time and space is considered. 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 from high-frequency data were studied previously by other researchers, who proved asymptotic normality of the MCEs. MCEs for unknown parameters of coordinate processes of the SPDEs in two space dimensions using thinned data with respect are obtained to space. By utilizing the MCEs, approximate coordinate processes of the SPDEs are constructed. Adaptive estimators are derived for the coefficient parameters of the SPDE using the approximate coordinate processes and thinned data concerning time. The adaptive estimators are proved to be asymptotically normal under some regular conditions. Simulation results of the proposed estimators are also presented.

E1026: **Parameter estimation for linear SPDEs: Discrete observations within the local approach**
Presenter: **Mathias Trabs**, Karlsruhe Institute of Technology, Germany
Co-authors: Randolf Altmeyer, Florian Hildebrandt

A linear SPDE on a bounded domain is studied driven by a stochastic noise process, which is white in time and possibly coloured in space. The aim is to bridge the gap between two popular observation schemes studied for statistics for SPDEs, namely, discrete observations and local measurements. To this end, the local measurements approach is extended to kernels of distribution type. In particular, the discrete Laplacian understood as a local measurement of distribution type, will allow for analyzing estimators based on discrete observations in arbitrary space dimensions.

EO063 Room 102 EMERGING STATISTICAL APPROACHES TO IMPROVE THE DEVELOPMENT OF CULTIVARS Chair: Hiroyoshi Iwata

E0788: **A Bayesian model for genomic prediction using metabolic networks**
Presenter: **Akio Onogi**, Ryukoku University, Japan

Genomic prediction is now an essential technique in animal and plant breeding, and it is also used for predicting disease risks in medicine. One of current research interests in genomic prediction would be how omics data can be used to improve prediction accuracy. A precedent work proposed a metabolic network-based method in biomass prediction of Arabidopsis. The method is based on flux balance analysis where production and consumption of all metabolites are assumed to be balanced. Although the idea is unique, the method consists of multiple steps that possibly degrade prediction accuracy. Here, I proposed a Bayesian model that integrates all steps and jointly infers all fluxes of reactions related to biomass production. The proposed model showed higher accuracies than methods compared both in simulated and real data. The findings support the previous idea that metabolic network information can be used for prediction.

E1187: **Bayesian optimization of genotype and environment interaction**
Presenter: **Hiroyoshi Iwata**, The University of Tokyo, Japan
Co-authors: Tai-Shen Chen, Chikashi Sato, Masanori Yamasaki, Chyon Hae Kim, Akira Abe, Hiroyuki Shimono

Stable food production requires the development of new varieties that can adapt to changes in the environment. Genomic selection can accelerate the development of such varieties. Note that since adaptation to future environments is a statistical extrapolation, the interaction between genotype and environment must be optimized, taking into account the uncertainty associated with the extrapolation. A Bayesian optimization (BO) method for genotype-environment interactions is proposed, using the analysis of historical Japanese rice breeding data as an example. Phenotypic variation caused by the combination of genotype and environment was modelled using a Gaussian process (GP), and simulations were performed using BO to search for the optimal combination of genotype and environment in the region to be extrapolated. The results showed that BO was able to detect the optimal combination more efficiently than point prediction in regions with high extrapolation. In addition, the optimal BO criterion was found to differ depending on the region of interest. In summary, it is found that the approach using modelling by GP and BO was effective in searching for the optimal combination of genotype and environment.

E1190: **Development of a novel GWAS method to detect QTL effects interacting with discrete and continuous genetic architecture**
Presenter: **Kosuke Hamazaki**, Center for Advanced Intelligence Project (AIP), RIKEN, Japan
Co-authors: Hiroyoshi Iwata, Tristan Mary-Huard

GWAS (Genome-Wide Association Study) aims at detecting candidate genes (QTL) associated with a target trait via statistical testing. Since a classical GWAS starts with the constitution of a panel of individuals, usually gathered from different populations, many methods have been proposed to control the false positive error rate in large datasets with a strong population structure. However, most methods assume the same QTL effect across populations, which is not always true in the natural biological process. A method has been previously proposed to consider population-specific QTL effects by testing marker effects in each population separately with prior information on population membership for each individual. However, this information on the population structure may only sometimes be available, and sometimes the population structure is more continuous than discrete, in which case the previous methodology cannot be applied. The proposed novel method does not require prior knowledge of the population structure. In the proposed models, we explicitly include an interaction term between an SNP/haplotype block of interest and the genetic background in the conventional SNP-based/haplotype block-based GWAS model. The proposed SNPxGB and HBxGB models can be justified because they can well consider the interaction between the QTLs and the discrete/continuous genetic architecture.

E1219: **Quantitative genetic analysis of metabolites in rice**
Presenter: **Gota Morota**, Virginia Polytechnic Institute and State University, United States

The asymmetric increase in average nighttime temperatures relative to the increase in average daytime temperatures due to climate change is decreasing grain yield and quality in rice. Therefore, a better genome-level understanding of the impact of higher night temperature stress on the weight of individual grains is essential for the future development of more resilient rice. The utility metabolites and single-nucleotide polymorphisms (SNPs) were investigated to predict grain length, width, and perimeter phenotypes using a rice diversity panel. Best linear unbiased prediction and BayesC showed greater metabolic prediction performance than machine learning models for grain-size phenotypes. The metabolic prediction was most effective for grain width, resulting in the highest prediction performance. Genomic prediction performed better than metabolic prediction. Integrating metabolites and genomics simultaneously in a prediction model slightly improved prediction performance. A difference in prediction between the control and HNT conditions is not observed. Several metabolites were identified as auxiliary phenotypes that could be used to enhance the multi-trait genomic prediction of grain-size phenotypes. The results showed that, in addition to SNPs, metabolites collected from grains offer rich information to perform regression modelling of grain-size-related phenotypes in rice.

E1221: **Comparing artificial-intelligence techniques with parametric prediction models for predicting soybean traits**
Presenter: **Reka Howard**, University of Nebraska - Lincoln, United States
Co-authors: Susweta Ray, Diego Jarquin

Soybean is a significant source of protein and oil and is also widely used as animal feed. Thus, developing lines that are superior in terms of yield, protein, and oil content is important to feed the ever-growing population. As opposed to high-cost phenotyping, genotyping is both cost and time efficient for breeders, thus enabling the potential success of genomic prediction techniques. A conventional GP method (genomic best linear unbiased predictor [GBLUP]), a kernel method (Gaussian kernel [GK]), an artificial intelligence (AI) method (deep learning [DL]), and a hybrid method that corresponds to the emulation of a DL model using a kernel method (an arc-cosine kernel [AK]) in terms of their prediction accuracies for predicting grain yield, oil, and protein using data from the soybean nested association mapping experiment are compared. The relative performance of the four methods varied with the response variable and whether the model included the genotype-by-environmental interaction (GE) effects or not. The GBLUP consistently showed better performances, whereas GK and AK followed a similar pattern to GBLUP, and DL performed slightly worse than the other three methods in most of the cases; however, this may also be attributed to suboptimal hyperparameters. The DL method performed particularly worse than the other three methods in the presence of the GE effects models.

EO247 Room 503 INFERENCE AND PREDICTION IN BAYESIAN NONPARAMETRICS Chair: Marta Catalano

E0364: **Random measure priors in Bayesian frequency recovery from sketches**
Presenter: **Mario Beraha**, Universita di Torino, Italy

Given a lossy-compressed representation, or sketch, of data with values in a set of symbols, the frequency recovery problem considers estimating the empirical frequency of a new data point. This is a classical problem in computer science, with recent studies applying Bayesian nonparametric (BNPs) to develop learning-augmented versions of the popular count-min sketch (CMS) recovery algorithm. A novel BNP approach is presented to frequency recovery, which is not built from the CMS but relies on a sketch obtained by random hashing. Assuming data to be modelled as random samples from an unknown discrete distribution, endowed with a Poisson-Kingman (PK) prior, given the sketch, the posterior distribution of a symbol's empirical frequency is provided, with estimates being obtained as posterior mean functionals. The BNP approach is further developed to a traits formulation of the frequency recovery problem, not yet considered in the CMS literature, in which data belong to more than one symbol, referred to as traits, and exhibit nonnegative levels of associations with each feature By modelling data as a generalized Indian buffet process, the posterior distribution of the empirical frequency level of a trait is provided, given a sketch obtained by random hashing. Some applications are presented for the Poisson and Bernoulli distribution for the levels of associations.

E0985: **Bayesian nonparametric inference by means of stick-breaking priors with dependent lenght variables**
Presenter: **Maria Fernanda Gil-Leyva Villa**, IIMAS,UNAM, Mexico
Co-authors: Ramses Mena, Pierpaolo De Blasi

The general classes of exchangeable stick-breaking processes (ESB) and Markov stick-breaking processes (MSB) are studied. In particular, the aim is to show how other well-known random probability measures, such as Dirichlet, Geometric and Pitman-Yor processes, can be recovered through ESBs and MSBs. The objective is to explain how to implement mixture models with an ESB or an MSB mixing priors by means of a novel Gibbs sampler method, and the performance of the models in clustering and density estimation is evaluated.

E0687: **Bayesian analysis for functional ANOVA model**
Presenter: **Yongdai Kim**, Seoul National University, Korea, South

The functional ANOVA model is a useful tool for constructing an interpretable prediction model. While there are several frequentist procedures to estimate the components in the functional ANOVA model (i.e. MARS, Splines), Bayesian procedures focus mainly on the generalized additive model (GAM), which is the simplest one among functional ANOVA models due to computational difficulties. A computationally efficient Bayesian procedure is proposed to infer components in the functional ANOVA model. The algorithm, called ANOVA-BART, is a modification of BART (Bayesian Additive Regression Tree), a well-known Bayesian procedure for estimating high-dimensional regression models. BART combines many baseline trees (simple trees) to infer the final prediction model. Even though BART is very good at prediction, its interpretation is not easy. To improve the interpretability of BART, the sets of baseline trees are constructed corresponding to each component of the functional ANOVA model, put prior to each set of baseline trees, and an MCMC algorithm is developed to search good linear combinations of baseline trees for each component.

E0778: **Nonparametric priors with fixed mean distributions**
Presenter: **Antonio Lijoi**, Bocconi University, Italy
Co-authors: Francesco Gaffi, Igor Pruenster

Linear functionals, or means, of discrete random probability measures, are natural probabilistic objects, and the investigation of their properties has a long and rich history. They appear in several areas of mathematics, including statistics, combinatorics, special functions, excursions of stochastic processes and financial mathematics, among others. Most contributions have aimed at determining their distribution starting from a fully specified random probability. The inverse problem is addressed: the identification of the base measure of a discrete random probability measure yielding a specific mean distribution. Besides its theoretical interest, this is of practical relevance to Bayesian nonparametric inference, where the law of a random probability measure acts as a prior distribution. Indeed, it is more often the case where pre-experimental information is available about a finite-dimensional projection of the data-generating distribution, such as the mean, rather than about an infinite-dimensional parameter. Results concerning the Dirichlet process, the normalized stable process and the Pitman-Yor process wil be displayed. They are further extended to nonparametric mixture models that are widely used for density estimation and clustering.

E0367: **Clustering consistency with Dirichlet process mixtures**
Presenter: **Giovanni Rebaudo**, University of Turin and Collegio Carlo Alberto, Italy
Co-authors: Filippo Ascolani, Antonio Lijoi, Giacomo Zanella

Dirichlet process mixtures are flexible nonparametric models particularly suited to density estimation and probabilistic clustering. The posterior distribution induced is studied by Dirichlet process mixtures as the sample size increases and, more specifically, focuses on consistency for the unknown number of clusters when the observed data are generated from a finite mixture. Crucially, the situation is considered where a prior is placed on the concentration parameter of the underlying Dirichlet process. Previous findings in the literature suggest that Dirichlet process mixtures are typically inconsistent for the number of clusters if the concentration parameter is fixed and data come from a finite mixture. It is shown that consistency for the number of clusters can be achieved if the concentration parameter is adapted in a fully Bayesian way, as commonly done in practice. The results are derived for data from a class of finite mixtures, with mild assumptions on the prior for the concentration parameter and for various choices of likelihood kernels for the mixture.

EO176 Room 506 MATHEMATICS OF DATA SCIENCE Chair: Dingxuan Zhou

E0966: **SStaGCN: Simplified stacking based graph convolutional networks**
Presenter: **Jia Cai**, Guangdong University of Finance and Economics, China

Graph convolutional network (GCN) is a powerful model studied broadly in various graph structural data learning tasks. However, designing GCN models to mitigate the over-smoothing phenomenon is still a crucial issue to be investigated. A novel Simplified Stacking is proposed based on GCN (SStaGCN) by utilizing stacking ideas, aggrehich is a general adaptive framework for tackling distinct types of structural graph data. Specifically, we first use the base models of stacking to extract the node features in the graph. Subsequently, aggregation methods such as mean, attention and voting techniques are employed to enhance the ability of feature extraction further. Thereafter, the node features are considered as inputs and fed into the vanilla GCN model. Furthermore, theoretical generalization bound analysis of the proposed model is explicitly given. Extensive experiments on 3 public citation networks and another 3 heterogeneous tabular data demonstrate the effectiveness and efficiency of the proposed approach over several state-of-the-art GCNs. Notably, the proposed SStaGCN can efficiently mitigate the over-smoothing problem of GCNs.

E1004: **Value-gradient based formulation of optimal control problem and machine learning algorithm**
Presenter: **Xiang Zhou**, City University of Hong Kong, Hong Kong

The optimal control problem is typically solved by first finding the value function through the Hamilton-Jacobi equation (HJE) and then taking the minimizer of the Hamiltonian to obtain the control. Instead of focusing on the value function, a new formulation is proposed for the gradient of the value function (value-gradient) as a decoupled system of partial differential equations in the context of a continuous-time deterministic discounted optimal control problem. This is similar to differential learning, but the derivation is simple and based on the Eulerian viewpoint, not from the underlying dynamics directly. An efficient iterative scheme is developed for this system of equations in parallel by utilizing the properties that share the same characteristic curves as the HJE for the value function. For the theoretical part, it is proven that this iterative scheme converges linearly. The characteristic line method is combined with machine learning techniques for the numerical method. Experimental results demonstrate that this new method not only significantly increases the accuracy but also improves the efficiency and robustness of the numerical estimates, particularly with less amount of characteristics data or fewer training steps.

E1061: **Lifting the veil: Unlocking the power of depth in Q-learning**
Presenter: **Shao-Bo Lin**, Xi'an Jiaotong University, China

With the help of massive data and rich computational resources, deep Q-learning has been widely used and received great success in numerous applications, including recommender systems, games and robotic manipulation. Compared with avid research activities in practice, there is a lack of solid theoretical verifications and interpretability for the success of deep Q-learning, making it a little bit mystery. The aim is to discuss the power of depth in deep Q-learning. In the learning theory framework, it is rigorously proven that deep Q-learning outperforms the traditional one by showing its good generalization error bound. The results show that the main reason for the success of deep Q-learning is due to the excellent performance of deep neural networks (deep nets) in capturing special rewards properties, such as the spatially sparse and piecewise constant, rather than due to their large capacities. In particular, answers are provided to questions about why and when deep Q-learning performs better than traditional ones and how the generalization capability of deep Q-learning.

E1056: **Theory of structured deep neural networks**
Presenter: **Dingxuan Zhou**, University of Sydney, Australia

Deep learning based on deep neural networks with network architectures has been powerful in practical applications but is less understood theoretically. The network structures give essential difficulty. An important family of structured deep neural networks is deep convolutional neural networks with convolutional structures. The convolutional architecture is key for computational efficiency but raises scientific challenges. A mathematical theory of approximating and learning functions or operators by deep structured deep neural networks is described.

E1062: **Classification with deep neural networks**
Presenter: **Lei Shi**, Fudan University, China
Co-authors: Zihan Zhang, Dingxuan Zhou

Classification with deep neural networks (DNNs) has made impressive advancements in various learning tasks. Due to the unboundedness of the target function, generalization analysis for DNN classifiers with logistic loss remains scarce. Recent progress in establishing a unified framework of generalization analysis for both bounded and unbounded target functions is reported. The analysis is based on a novel oracle-type inequality, which enables us to deal with the boundedness restriction of the target function. In particular, for logistic classifiers trained by deep, fully connected neural networks, the optimal convergence rates are obtained only by requiring the Hölder smoothness of the conditional probability. Under certain circumstances, such as when decision boundaries are smooth and the two classes are separable, the derived convergence rates can be independent of the input dimension.

EO036 Room 603 CURRENT DEVELOPMENTS IN QUANTITATIVE FINANCE Chair: Rogemar Mamon

E0608: **An enhanced neural network approach for agricultural index insurance design**
Presenter: **Heng Xiong**, Wuhan University, China
Co-authors: Runqiu Xu, Wenjun Jiang, Rogemar Mamon

Determining the proper payoff function and corresponding premium is essential to agricultural index insurance design. However, prevailing index insurance contracts encounter high basis risk and thus charge an unwilling premium. First, a non-linear payoff function structure utilizing a neural network with a multivariate weighted premium principle (MWPP) is proposed. Then the payoff function with a utility maximization problem is optimized to form the index insurance contract. Compared to traditional linear payoff schemes, neural networks capture non-linearity between utility-maximized payoff and indices, while MWPP provides viable premiums that offer more fair and accurate premium rates. Further, our approach is examined using China's grid-cell rice yield, weather, and soil data. Empirical results present that the proposed method reduces basis risk and improves insurers' utility with the more actual payoff and lower premiums.

E0911: **Optimal commissions and subscriptions in mutual aid platforms**
Presenter: **Yixing Zhao**, Guangdong University of Foreign Studies, China

An operation mechanism for mutual aid platforms is investigated to develop sustainably and profitably. A mutual aid platform is an online risk-sharing platform for risk-heterogeneous participants, and the platform extracts revenues by charging participants commission and subscription fees. A modelling framework is proposed to identify the optimal commissions and subscriptions for mutual aid platforms. Participants are divided into different types based on their loss probabilities and values derived from the platform. It is presented how these commissions and subscriptions should be set in a mutual aid plan to maximize the platform's revenues. The analysis emphasized the importance of accounting for risk heterogeneity in mutual aid platforms. Specifically, different types of participants should be charged different commissions/subscriptions depending on their loss probabilities and values on the platform. Participants' shared costs should be determined based on their loss probabilities. Adverse selection occurs on the platform if participants with different risks pay the same shared costs. The results also show that the platform's maximum revenue will be lower if the platform charges the same fee to all participants. The numerical results of a practical example illustrate that the optimal commission/subscription scheme and risk-sharing rule result in considerable improvements in platform revenue over the current scheme implemented by the platform.

E1074: **Pricing formulas for perpetual American options with general payoffs**
Presenter: **Mariano Rodrigo**, University of Wollongong, Australia

An American option gives the holder the right, but not the obligation, to buy/sell an underlying asset from/to the writer at an agreed strike price at any time on or before the expiry date. Options are mainly used for speculation and hedging. One of the attractions of options is that they can be used to construct a wide range of trading strategies characterised by different payoff functions. The pricing of perpetual American options with general payoffs is considered, where the perpetual American call and put are special cases. Four broad classes of payoff functions are identified for which analytical pricing formulas can be derived by utilising a Mellin transform technique and an optimisation procedure. Free boundary problems with one or two boundaries are obtained depending on the class of payoff functions considered. Illustrative examples are provided and benchmarked numerically with the binomial method. The characterisation of different payoffs for perpetual American options considered here will be instrumental in identifying and pricing new free boundary problems for (non-perpetual) American-style financial derivatives.

E1101: **Bayesian nonlinear expectation for time series modelling and its application to Bitcoin**
Presenter: **Tak Kuen Siu**, Macquarie University, Australia

A two-stage approach to parametric nonlinear time series modelling in discrete time is proposed with the objective of incorporating uncertainty in the conditional mean and volatility. A reference time series model is specified and estimated in the first stage. In the second stage, Bayesian nonlinear expectations are introduced to incorporate model uncertainty in prediction by specifying a family of alternative models. The construction of Bayesian nonlinear expectations for prediction is based on closed-form Bayesian credible intervals evaluated using conjugate priors and residuals of the estimated reference model. Using real Bitcoin data, including some periods of Covid 19, the proposed method is applied to forecast Bitcoin returns and evaluate Bitcoin risks under three major parametric nonlinear time series models, namely the self-exciting threshold autoregressive model, the generalized autoregressive conditional heteroscedasticity model, the stochastic volatility model.

E1049: **A signal-processing approach in cyber risk valuation**
Presenter: **Rogemar Mamon**, University of Western Ontario, Canada

The cyber risk insurance market is rapidly developing due to the potentially huge losses from cyberattacks. We present such a framework for cyber risk modelling, wherein the cyberattacks occurrences and their inter-arrival and duration are captured by a regime-switching Markov model (RSMM). In this customised RSMM, the transition probabilities of the Markov chain are governed by another hidden Markov chain representing the various states of the cyber security environment. A self-calibrating mechanism is provided via filtering, and a cyber kill chain is built based on the stages of the cyberattack. With the aid of change of reference probability measures and the EM algorithm, the estimators for the transition matrix are derived. The main point of interest is the random losses from cyberattacks, which are assumed to follow a doubly truncated Pareto distribution. The Vasicek model is utilised to describe the interest rate process for the discounting of losses. The premium for a cyber security insurance contract is calculated via a simulated data set based on two pricing principles. The methodology featuring dynamic parameter estimation and flexible adjustments in modelling various risk factors widens the available pricing and cyber risk management tools.

EO026 Room 604 BAYESIAN COMPUTATION FOR COMPLEX MODELS Chair: David Nott

E0255: **Efficient variational approximations for state space models**
Presenter: **Ruben Laoiza Maya**, Monash University, Australia
Co-authors: Didier Nibbering

Variational Bayes methods are a scalable estimation approach for many complex state space models. However, existing methods exhibit a trade-off between accurate estimation and computational efficiency. A variational approximation that mitigates this trade-off is proposed. This approximation is based on importance densities that have been proposed in the context of efficient importance sampling. By direct conditioning on the observed data, the proposed method produces an accurate approximation to the exact posterior distribution. Because the steps required for its calibration are computationally efficient, the approach is faster than existing variational Bayes methods. The proposed method can be applied to any state-space model that has a closed-form measurement density function and a state transition distribution that belongs to the exponential family of distributions. The method is illustrated in numerical experiments with stochastic volatility models and a macroeconomic empirical application using a high-dimensional state space model.

E0441: **Sampling with constraints**
Presenter: **Xin Tong**, National University of Singapore, Singapore

Sampling-based inference and learning techniques, especially Bayesian inference, provide an essential approach to handling uncertainty in machine learning (ML). As these techniques are increasingly used in daily life, it becomes essential to safeguard ML systems with various trustworthy-related constraints, such as fairness, safety, and interpretability. A family of constrained sampling algorithms which generalize Langevin Dynamics (LD) and Stein Variational Gradient Descent (SVGD) is proposed to incorporate a moment constraint or a level set specified by a general nonlinear function. By exploiting the gradient flow structure of LD and SVGD, algorithms are derived for handling constraints, including a primal-dual gradient approach and the constraint-controlled gradient descent approach. The continuous-time mean-field limit of these algorithms is investigated, and it is shown that they have O(1/t) convergence under mild conditions.

E0549: **Bayesian score calibration for approximate models**
Presenter: **Christopher Drovandi**, Queensland University of Technology, Australia
Co-authors: Joshua Bon, David Nott, David Warne

Scientists continue to develop increasingly complex mechanistic models to reflect their knowledge more realistically. Statistical inference using these models can be highly challenging since the corresponding likelihood function is often intractable, and model simulation may be computationally burdensome. Fortunately, in many of these situations, it is possible to adopt a surrogate model or approximate likelihood function. It may be convenient to base Bayesian inference directly on the surrogate, but this can result in bias and poor uncertainty quantification. Here a new method for adjusting approximate posterior samples is proposed to reduce bias and produce more accurate uncertainty quantification. This is done by optimising a transform of the approximate posterior that maximises a scoring rule. Our approach requires only a (fixed) small number of complex model simulations and is numerically stable. The good performance of the new method on several examples of increasing complexity is demonstrated.

E0550: **Semiparametric Bayesian two-stage meta-analysis for association between ambient temperature and new cases of COVID-19**
Presenter: **Dongu Han**, Korea University, Korea, South
Co-authors: Dongu Han, Kiljae Lee, Yeonseung Chung, Taeryon Choi

Two-stage meta-analysis has been a popular tool to investigate a short-term association between environmental exposure and a healthy response; in the first stage, a generalized linear model with distributed lag structure is typically fitted for each location and in the second stage, the location-specific association parameters estimated in the first stage are pooled to generate a combined estimate. A novel Bayesian approach for a two-stage meta-analysis is proposed, and an efficient MCMC algorithm and fast variational Bayes algorithms are developed as an alternative for each stage of the proposed model. Precisely, for the first stage, a new Bayesian distributed lag nonlinear model, which accommodates complex nonlinearities among time, covariate and lag, are proposed, and the model is estimated by utilizing non-conjugate variational message passing and importance sampling. A robust matrix-variate Dirichlet process mixture multivariate meta-regression is proposed for the second stage model, and a fast online variational Bayes approach is developed to estimate the model. The proposed methods are illustrated by applying them to study a short-term association between ambient temperature and new cases of COVID-19 in the United States.

E0962: **Fast Bayesian estimation of dynamic linear regression models for semi long memory time series**
Presenter: **Matias Quiroz**, University of Technology Sydney, Australia
Co-authors: Thomas Goodwin, Robert Kohn

Dynamic linear regression models forecast the values of a time series based on a linear combination of a set of exogenous time series while incorporating a time series process for the error term. This error process is often assumed to follow an autoregressive integrated moving average (ARIMA) model, or seasonal variants thereof, which are unable to capture a long-range dependency structure of the error process. A novel dynamic linear regression model that incorporates the long-range dependency feature of the errors is proposed, showing that it improves the model's forecasting ability. A Markov chain Monte Carlo method is developed to fit general dynamic linear regression models based on a frequency domain approach that enables fast approximate Bayesian inference for large datasets. It is demonstrated that the approximate algorithm is much faster than the traditional time domain approach based on the Kalman filter while retaining high accuracy.

EO089 Room 605 RECENT ADVANCES IN TIME SERIES MODELING Chair: George Michailidis

E0666: **Inference for nonstationary timeseries using optimal Gaussian approximation with explicit construction**
Presenter: **Sayar Karmakar**, University of Florida, United States

Inference problems for time series, such as curve estimation for time-varying models or testing for the existence of change-point, have garnered significant attention. However, these works are restricted to the limiting assumption of independence and/or stationarity at their best. The main obstacle is that the existing optimal Gaussian approximation results for nonstationary processes only provide an existential proof, and thus they are difficult to apply. A clear path is provided to construct such a Gaussian approximation. The proposed Gaussian approximation results encapsulate a very large class of nonstationary time series, obtain the optimal rate and yet have good applicability. Building on such a Gaussian approximation, theoretical results for changepoint detection and simultaneous inference in the presence of nonstationary errors are shown. The theoretical results with extensive simulation studies and some real data analyses are substantiated.

E0702: **Nonpivotal Granger causality tests based on vector autoregressive models**
Presenter: **Ying-Chao Hung**, National Taiwan University, Taiwan

Granger causality is a classical and important technique for measuring predictability from one group of time series to another by incorporating information on variables described by a vector autoregressive (VAR) model. Traditional methods for validating Granger causality are based on the Wald type tests, which may encounter the following implementation problems: (i) test statistic inflation due to singularity or near-singularity of the underlying covariance matrix; and (ii) infeasibility or huge computational cost for tuning parameter selection. An alternative procedure for testing Granger causality based on non-pivotal statistics is considered. The proposed testing method has a strong theoretical basis and does not require any calibration of tuning parameters. Further, an initial numerical investigation yields very competitive power values compared to the Wald-type tests. Finally, the purpose is to extend the proposed method and establish associated asymptotic theories for large orders (or infinite order) of vector autoregressive models.

E0804: **Inference on the change point under a high dimensional covariance shift**
Presenter: **Abhishek Kaul**, Washington State University, United States

The focus is on the problem of constructing asymptotically valid confidence intervals for the change point in a high-dimensional covariance shift setting. A novel estimator for the change point parameter is developed, and its asymptotic distribution under high dimensional scaling is obtained. First, it is established that the proposed estimator exhibits a sharp convergence rate. Further, the form of the asymptotic distributions under both a vanishing and a non-vanishing regime of the jump size is characterized. In the former case, it corresponds to the argmax of an asymmetric Brownian motion, while in the latter case to the argmax of an asymmetric random walk. The relationship between these distributions is then obtained, which allows the construction of regime (vanishing vs non-vanishing) adaptive confidence intervals. Easy-to-implement algorithms for the proposed methodology are developed, and their performance is illustrated on synthetic and real data sets.

E0850: **Influencer detection in market sectors via sparse network analysis**
Presenter: **Simon Trimborn**, University of Amsterdam, Netherlands
Co-authors: Kexin Zhang

When financial market participants expect that news relating to a company is representative of other companies within the same sector, then the performance of that company on the markets is expected to drive the other assets' performance as well. Such situations often arise during earning announcement season. A Sparse Network Model (SNM) is introduced to identify the influential assets within sectors. Usually, sectors comprise a large number of assets relating to the issue of high dimensionality. Naturally, not all assets within a sector are expected to impact the performance of others; hence often, a sparse underlying structure arises. Sparse estimation techniques are often applied to uncover such structures. When particular structures like groups or blocks are part of the network, then a tailored estimator commonly provides a more accurate estimation. As such, an estimator is introduced to detect influencers in asset networks. The methodology is flexible. As such, it extends various sparsity estimation techniques towards detecting influencers when they are present. The asymptotic properties of the estimator are studied, and its performance in extensive synthetic data experiments is validated. The impact of assets is studied on others within the sectors of the S&P100. The aim is to illustrate which companies are most influential for the sector and document the dynamics in the influencer structure over time.

E0895: **A general framework for network autoregressive processes**
Presenter: **George Michailidis**, University of Florida, United States

A general, flexible framework for Network Autoregressive Processes (NAR) is developed, wherein the response of each node in the network linearly depends on its past values, a prespecified linear combination of neighbouring nodes and a set of node-specific covariates. The corresponding coefficients are node-specific, and the framework can accommodate heavier than Gaussian errors with spatial-autoregressive, factor-based, or in specific settings, general covariance structures. A sufficient condition is provided that ensures the stability (stationarity) of the underlying NAR that is significantly weaker than its counterparts in previous work in the literature. Further, ordinary and (estimated) generalized least squares estimators are developed for both fixed and diverging numbers of network nodes and provide their ridge regularized counterparts that exhibit better performance in large network settings, together with their asymptotic distributions. Their asymptotic distributions are derived that can be used for testing various hypotheses of interest to practitioners. The issue of misspecifying the network connectivity and its impact on the aforementioned asymptotic distributions of the different NAR parameter estimators are also addressed. The framework is illustrated on both synthetic and real air pollution data.

EO076 Room 606 PROGRESS IN LEARNING AND MODELLING OF COMPLEX TIME SERIES AND SPATIAL DATA Chair: Zudi Lu

E0384: **Estimation of threshold dynamic regression for cross-sectional dependent panel time series**
Presenter: **Zudi Lu**, University of Southampton, United Kingdom
Co-authors: Lulu Wang, Maria Kyriakou

The estimation of a family of dynamic threshold models is studied for cross-sectional dependent panel time series. Although the idea of threshold popular in nonlinear analysis for time series has been extended to panel data under cross-sectional independence (CSI), the inference tools built under the CSI, cannot apply to such financial analysis of panel stocks owing to intrinsic cross-sectional dependence (CSD). An estimation of the problem has hence developed under the CSD, with the least squares coefficient estimators shown to be asymptotically normal at a convergence rate of root-nT (with n and T for the sample sizes in cross-section and in time), but their asymptotic variance matrix viably different from that under the CSI. The consistent estimator of the asymptotic variance matrix is further constructed under the CSD. The conditions ensuring the estimators of threshold parameters having a non-standard asymptotic distribution under the CSD are sought in a general setting with the threshold effects diminishing at varied rates in n and T. Monte Carol simulations on the performance under finite sample show that the variances of the estimators would be significantly underestimated, leading to spurious inference, if the panel's cross-sectional dependency were ignored or mistaken to be as the CSI. An empirical application to study the effect of the precipitation on the panel of stocks of FTSE100 confirms that the threshold method facilitates an effective financial analysis.

E0479: **Generalising dynamic semiparametric averaging forecasting for time series with discrete-valued response**
Presenter: **Fangsheng Ge**, University of Southampton, United Kingdom
Co-authors: Rong Peng, Zudi Lu

The aim is to explore how to utilise the useful high-dimensional lagged information for dynamic forecasting of time series data with a discrete-valued response. The approach will generalise the existing flexible semiparametric marginal regression model averaging (MARMA) forecasting, a useful data-driven method designed for nonlinear forecasting of continuous-valued time series by the least squares averaging. A generalised MARMA (GMARMA) procedure has been suggested under a general time series exponential family of distributions, which flexibly accommodates nonlinear forecasting of discrete-valued response. Further, it allows lagged effects, including discrete-valued information for forecasting. A conditional likelihood model averaging method, instead of the least squares, is developed to estimate the average weights in the GMARMA under a beta-mixing time series data generating process with established asymptotic normality. Furthermore, an adaptively penalised GMARMA (PGMARMA) is suggested to select the important variables for improved forecasting. The oracle properties of the PGMARMA weights are established as if the true non-zero weights were known. These procedures are further supported by Monte Carlo simulations and empirical applications to forecasting the FTSE 100 index market moving direction and the UK road casualty data, which outperform many popular machine learning tools.

E0524: **Robust reduced rank estimation for low-rank vector AR models**
Presenter: **Fumiya Akashi**, University of Tokyo, Japan

The estimation problem of vector autoregressive models with possibly infinite variance error processes is considered. A general low-rank structure for the coefficient matrices is assumed, and an estimation procedure based on the multivariate median is proposed. The reduced rank estimation algorithm is also provided, and the proposed estimator is shown to improve the efficiency of the classical least-distance estimator in some sense. Some simulation experiments illustrate the finite sample performance of the proposed method.

E0658: **Robust inference on infinite and growing dimensional time series regression**
Presenter: **Abhimanyu Gupta**, University of Essex, United Kingdom
Co-authors: Myung Hwan Seo

A class of tests for time series models, such as multiple regression with growing dimension, infinite-order autoregression and nonparametric sieve regression, are developed. Examples include the Chow test and general linear restriction tests of growing rank p. Employing such increasing p asymptotics, a new scale correction is introduced to conventional test statistics, which accounts for a high-order long-run variance (HLV) that emerges as p grows with sample size. A bias correction via a null-imposed bootstrap is also proposed to alleviate finite sample bias without sacrificing power. A simulation study shows the importance of robustifying testing procedures against the HLV even when p is moderate. The tests are illustrated with an application to the oil regressions in Hamilton (2003).

E0731: **Predictive models for time series by deep structured learning**
Presenter: **Shubin Wu**, University of Southampton, United Kingdom
Co-authors: Zudi Lu

Empirical studies on time series forecasting with deep neural networks have become widespread, but there is still no clear conclusion on which method works well, posing significant mysteries. The impact of the structure of a deep learning model on prediction, specifically examining the asymptotic properties of a multiple-layer neural network parametric autoregression under time series beta-mixing conditions, is investigated. An hourly wind electricity production time series dataset is used to compare three popular models: autoregressive multiple layer perceptron deep learning (AR-MLP), long-short term memory neural network (LSTM), and residual neural network (ResNet). The accuracy of these models is measured in terms of mean square error (MSE) for the predictions, and an interpretable model that reasonably fits the hidden layers of the neural network is explored through a reasonable selection of the AR lag order for the input with an equal number of neurons in each hidden layer and the number of hidden layers. The results show that the AR-MLP model achieves high prediction accuracy and efficiency after considering validation set MSE, test MSE, and training time, suggesting that a neural network model with a relatively simple structure can potentially provide good prediction performance.

EO020 Room 702 MULTIVARIATE PROBLEMS FOR STRUCTURED DEPENDENT DATA I Chair: Matus Maciak

E0418: **Semi-continuous time series with volatility clustering**
Presenter: **Sarka Hudecova**, Charles University, Prague, Czech Republic

Time series that contain a non-negligible portion of possibly dependent zeros, whereas the remaining observations are positive, are considered. They are treated as GARCH processes consisting of non-negative values. Such models find application in various fields and are, to some extent, related to multiplicative error models (MEM). The aim lies in the estimation of the omnibus model parameters while taking into account the semi-continuous distribution. The hurdle distribution, together with dependent zeros, causes the classical GARCH estimation techniques to fail, so two different estimation approaches are proposed. The resulting two quasi-likelihood estimators are shown to be strongly consistent and asymptotically normal. The empirical properties are illustrated in a Monte Carlo simulation study, demonstrating the computational efficiency of the methods employed. The developed techniques are presented through an actuarial problem concerning sparse insurance claims.

E0845: **Self-normalising change-point detection procedures for high-dimensional data**
Presenter: **Charl Pretorius**, North-West University, South Africa
Co-authors: Heinrich Roodt

Nonparametric CUSUM-based test criteria are presented for detecting changes in the means of high-dimensional panel data. The test statistics are shown to be self-normalising in the sense that their null distributions are asymptotically pivotal, even in the presence of weak serial dependence. Hence, unlike many existing procedures, the new tests have the practical advantage of not requiring estimation of the long-run variance of the error component in each panel and therefore eliminate the reliance on choosing bandwidth parameters. This allows for the tabulation of general asymptotic critical values, which may readily be used in applications, including in situations where the true underlying data-generating process is unknown. The results are further generalised to the case where the panels are allowed to depend on unobserved common factors. Numerical results are presented, which show that the new tests are level-respecting under the null hypothesis for large enough samples and under weak time-dependence. An application to real data is discussed, and it is shown that the tests perform well when the innovations follow popular financial time series models such as ARMA and GARCH models.

E0291: **Regime changes, I-phenomena, and unsupervised learning**
Presenter: **Michal Pesta**, Charles University, Czech Republic
Co-authors: Marie Huskova

The purpose is to deal with a situation such that every occurrence of a phenomenon can cause several related events, and each event contributes to a different univariate counting process. Therefore, a collection of these dependent point processes forms a flexible multivariate counting process, where neither stationarity nor independence of interarrival times of the marginal processes is assumed. The main aim is to detect a structural break of some phenomena'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. The asymptotic behaviour of the test statistic under the null hypothesis and also under the alternatives are investigated. Bootstrap add-on is proposed to overcome the computational curse of dimensionality and avoid nuisance parameters. The validity of the resampling technique is proved. A changepoint estimator is introduced as a by-product, and its consistency is provided. Multiple changepoints' detections are designed. The empirical properties are illustrated in a simulation study. The completely data-driven detection procedure is presented through an actuarial problem concerning claims from various insurance lines of business.

E0768: **Hockey is a cruel game: Empirical Bayes woes in predicting productivity of hockey players**
Presenter: **Ivan Mizera**, University of Alberta, Canada

A prominent domain of illustrations/applications of the potential of empirical Bayes methodology is the prediction of the overall performance of players in sports from rather scant data, achievable by borrowing strength from their peers. The existing instances of such data-analytic ventures, from illuminating examples to comprehensive studies, considered rather binomial responses of hits-and-misses, like bats in baseball or penalties in basketball. The Poisson modelling of point events, also featured in very early empirical Bayes pursuits mainly in the context of actuarial data, was not that much exposure in the sports context; however, the scheme of evaluating productivity in hockey via goals and assists leads exactly to this setting. Relevant parametric and nonparametric approaches of the empirical Bayes methodology therein, adapted to and comparing the specifics of the game of ice hockey to the actuarial situations, are confronted; among other aspects, ramifications like stratification and including covariates are considered. Some preliminary evaluations of the methodology on the data from the NHL season 2018/2019 are presented, and some directions for future reflection and exploration are discussed.

E0290: **Testing exchangeability of multivariate distributions**
Presenter: **Jan Kalina**, The Czech Academy of Sciences, Institute of Information Theory and Automation, Czech Republic

Although there have been a number of available tests of bivariate exchangeability, i.e. bivariate symmetry for bivariate distributions, the literature is void of tests on whether a multivariate distribution with more than two dimensions is exchangeable or not. Multivariate permutation tests of exchangeability of multivariate distributions are proposed, which are based on the non-parametric combination methodology, i.e. on combining non-parametric bivariate exchangeability tests. Numerical experiments on real as well as simulated multivariate data with more than two dimensions are presented here. The multivariate permutation test turns out to be typically more powerful than a bivariate exchangeability test performed only over a single pair of variables and also more suitable compared to tests exploiting the approaches of Benjamini-Yekutieli or Bonferroni.

EO122 Room 703 SPATIAL AND SPATIO-TEMPORAL METHODS Chair: Hsin-Cheng Huang

E0535: **Changes in extreme rainfall in Taiwan**
Presenter: **Nan-Jung Hsu**, National Tsing Hua University, Taiwan
Co-authors: Cheng-Ching Lin

Global warming is a critical issue for the ecosystem and environment on Earth. Many studies have found evidence that global warming has impacts on more frequent and serious climate conditions. The purpose is to examine the impact of global warming on extreme rainfalls in Taiwan using extreme value analysis. The annual maximum of daily rainfall is characterized using a PGEV model. In particular, both the intensity and the frequency parameters are assumed to have a spatially-varying regression relationship to the global temperature anomaly. The model parameters are estimated using the maximum likelihood. Our analysis found strong statistical evidence that the intensity of extreme rainfall in Taiwan becomes more severe as the temperature increases, especially in southern and central regions. Moreover, the extreme rainfall frequency also increases as the temperature increases, especially in southern and northern regions. Further, the return levels of annual extreme daily rainfall are evaluated under four hypothetical scenarios of temperature arising from 0.5C to 3C. Even if the global temperature only rises 1C, the frequency of seeing Taiwan's current 20-year return level will be doubled.

E0506: **Nonstationary Gaussian scale mixtures for spatial extremes**
Presenter: **Yen-Shiu Chin**, National Tsing Hua University, Taiwan
Co-authors: Nan-Jung Hsu, Hsin-Cheng Huang

Modelling spatial extremes are considered using Gaussian scale mixture models. This type of model includes asymptotic spatial dependence and asymptotic spatial independence scenarios according to the tail decay rates for their random scales. However, existing Gaussian scale mixture models are restricted to be stationary (or even isotropic), which is inappropriate for many spatial extreme applications. A nonstationary Gaussian scale mixture model using a basis-function approach is proposed. The maximum likelihood (ML) method is employed to estimate the model parameters and the ML estimators are computed via an expectation-maximization algorithm. The performance of the model is compared with stationary Gaussian scale mixture models through several simulated examples.

E0523: **Spatio-temporal analysis of dependent risk with an application to cyberattacks data**
Presenter: **Songhyun Kim**, Seoul National University, Korea, South
Co-authors: Chae Young Lim, Yeonwoo Rho

Cybersecurity is an important issue, given the increasing risks due to cyberattacks in many areas. Cyberattacks could result in huge losses such as data breaches, failures in the control systems of infrastructures, physical damages in manufacturing industries, etc. As a result, cybersecurity-related research has grown rapidly for in-depth analysis. The main interest is to understand the correlated nature of cyberattack data. To understand such characteristics, a spatial-temporal model is proposed for the host-wisely aggregated cyberattack data by incorporating the characteristics of the attackers. A new dissimilarity measure as a proxy of spatial distance is developed to be integrated into the model. The proposed model can be considered a spatial extension of the GARCH model. The estimation is carried out using a Bayesian approach, which is demonstrated to work well in simulations. The proposed model is applied to publicly available honeypot data after the data are divided by selected features of the attackers via clustering. The estimated model parameters vary by groups of attackers, which was not revealed by modelling the entire dataset.

E0559: **Contiguous segmentation of second-order non-stationary spatial processes**
Presenter: **ShengLi Tzeng**, National Sun Yat-sen University, Taiwan
Co-authors: Bo-Yu Chen, Hsin-Cheng Huang

Geostatistics, as a subfield of statistics analyzing and modelling spatial data, commonly assumes the process of interest to be stationary. Although this assumption is typically satisfied in small areas, it may not be appropriate for larger areas or more complex spatial phenomena. The nonstationary problems where the spatial covariance changes over the spatial domain are considered. To address this issue, a novel method for testing stationarity is proposed. Our method utilizes robust local estimates of spatial covariances to derive a test statistic. The data locations are then clustered using Voronoi tessellations. Furthermore, in cases where the stationary assumption is violated, a method is provided for identifying nonstationary features by partitioning the region into more homogeneous and close-to-stationary subregions. The optimal number of partitions can be determined using the Bayesian information criterion, ensuring our method is accurate and efficient. Notably, our proposed method applies to irregularly spaced data, making it a versatile tool for exploring a wide range of spatial data sets.

E0538: **Nonstationary spatial modeling, estimation, and prediction using a divide-and-conquer approach**
Presenter: **Hsin-Cheng Huang**, Academia Sinica, Taiwan
Co-authors: Chun-Shu Chen, Yung-Huei Chiou

Spatial data over a large domain generally shows nonstationary spatial covariance characteristics. However, estimating a nonstationary covariance function from a single realization of data is challenging, and the computation of the optimal spatial prediction is intractable when the dataset is massive. Initially, a method is proposed for visualizing nonstationary covariance structures, and a statistical test is introduced for spatial stationarity. Upon detection of nonstationarity, a segmentation technique is proposed that decomposes the spatial domain into K subregions wherein the process is approximately stationary. A stationary process is also considered for each of these K subregions. Subsequently, a novel nonstationary model that employs a linear combination of these processes with spatially varying weights is developed. Contrary to independent stationary models, our approach treats the K stationary processes as interdependent and represents them using a multivariate Matern covariance model. The proposed nonstationary model showcases flexibility, morphing into a globally stationary process when all stationary components exhibit a shared spatial covariance structure. Finally, a divide-and-conquer strategy for fast spatial prediction is proposed. The effectiveness of our approach is demonstrated through numerical experiments.

EO312 Room 704 STATISTICAL DESIGN AND MODELING FOR LIFE AND CLIMATE SCIENCES Chair: Andreas Futschik

E0590: **A convex approach to optimum design of experiments with correlated observations**
Presenter: **Werner Mueller**, Johannes Kepler University Linz, Austria
Co-authors: Andrej Pazman, Markus Hainy

The optimal design of experiments for correlated processes is an increasingly relevant and active research topic. Present methods have restricted possibilities to judge their quality. To fill this gap, the virtual noise approach is complemented with a convex formulation leading to an equivalence theorem comparable to the uncorrelated case and to an algorithm giving an upper performance bound against which alternative design methods can be judged. Moreover, a method for generating exact designs follows naturally. Estimation problems are exclusively considered on a finite design space with a fixed number of elements. A comparison of some classical examples from the literature as well as a real application is provided.

E0564: **Genetic adaptations in the population history of Arabidopsis thaliana**
Presenter: **Hirohisa Kishino**, Chuo University, Japan
Co-authors: Reiichiro Nakamichi, Shuichi Kitada

During the course of its range expansion or due to an environmental change, the population of Arabidopsis thaliana encountered unexperienced biotic and abiotic stresses. The dynamics of the allele frequencies at the 89,786 QTLs of 248 traits and 21,914 eQTLs of 2,879 genes identified 650 phenotypic adaptations (p-adaptations) and 3,925 gene expression-adaptations (e-adaptations) (FDR=0.05) were analyzed. The population accomplished large-scale p-adaptations and e-adaptations along four lineages, the eastward migration to Central Asia, northward migration to Scandinavia, migration to Azerbaijan, and migration into the United States. Extremely cold winters and short summers extended seed dormancy and expanded the root system architecture. Low temperatures lengthened the growth periods, and low light intensity necessitated increased chloroplast activity. The subtropical and wet environment enhanced phytohormone signalling pathways, responding to the biotic and abiotic stresses. Being exposed to heavy metals, the alleles underlying lower uptake from the soil, lower growth rate, lower resistance to bacteria, and higher expression of photosynthetic genes were selected. In total, 34,885 changes in allele frequencies were identified beyond the level of genetic drift (FDR=0.05). The database of QTLs and eQTLs combined with climatic information enabled a knowledge-based population genomic analysis, providing a clue for understanding the complex history of biological adaptation.

E0546: **A new approach for estimating the largest mean in a Gaussian mixture model with applications in population genetics**
Presenter: **Andreas Futschik**, JKU Linz, Austria

A new method is proposed to estimate the mixture component with the largest mean parameter when the data come from a Gaussian mixture model. Some properties of the method are discussed, and it is shown that it has advantages compared to classical approaches of inference, such as the EM algorithm, when there are many components. The method relies on inference for the truncated normal distribution. Our motivating application comes from population genetics, where the effective population size N_e is an important parameter when specifying null models. It first is explained how N_e has usually been estimated. Then it is shown how our proposed method may be used to identify the neutral N_e.

E0749: **Haplotype reconstruction via Bayesian linear models with unknown design**
Presenter: **Yuexuan Wang**, Johannes Kepler University Linz, Austria

The topic is the reconstruction of the unknown matrices S and ω for the multivariate linear model $Y = S\omega + \varepsilon$ under the assumption of binary entries $s_{ij} \in \{0,1\}$ for S and ω is a weight matrix. While a frequentist method has recently been proposed for this purpose, a Bayesian approach also seems desirable. In contrast to the point estimates provided by this frequentist method, our proposed hierarchical model delivers a posterior that permits quantifying uncertainty. Since matching permutations in both S and ω lead to the same reconstruction $S\omega$, an order-preserving shrinkage prior is introduced to establish identifiability concerning permutations. For inference, a blocked Metropolis-Hastings is introduced within the Gibbs sampling scheme to sample from the hierarchical model enforcing all constraints.

E0612: **Robustifying an exact test for heteroscedasticity in a two-way layout in variety frost trials with a covariate**
Presenter: **Brenton Clarke**, Murdoch University, Australia
Co-authors: Angelika Pilkington, Dean Diepeveen

Testing for heteroscedasticity in two-way layouts is often fraught with difficulty. Established tests are often hampered by the test statistic being based on asymptotics when the number of replications per cell can be as low as one. Moreover, as in various frost trials, there is the potential to include one or more covariates. The simple form of heteroscedasticity considered is dichotomous. For example, a group of varieties that may have larger response values can be tested to see if they have increased variance. The considered test is an F-test based on vectors of transformed residuals from the linear model. The exact F-test is shown to be more powerful than the asymptotic likelihood ratio test based on restricted or residual maximum likelihood (REML) available in statistical packages. However, it is, as are both these tests, not robust. Pearson noted the F-test's lack of robustness early last century. The size and power of a robustified F-test based on the Levene statistic formulated from the vectors of transformed residuals without and also with contamination are examined. The implementation of our test is illustrated using wheat data used in various frost trials carried out in Western Australia when a covariate is necessary for the modelling.

EO015 Room 705 RECENT ADVANCES IN HIGH-DIMENSIONAL DATA ANALYSIS Chair: Kazuyoshi Yata

E0786: **Asymptotic behaviors of k-means under high dimensional settings**
Presenter: **Kento Egashira**, Tokyo University of Science, Japan
Co-authors: Kazuyoshi Yata, Makoto Aoshima

While k-means has been approved as a useful methodology for analysing gene expression microarray data on behalf of high-dimensional, low-sample-size (HDLSS), k-means is not sufficiently studied theoretically under high-dimensional settings. The asymptotic properties of k-means are proved under mild and practical settings for HDLSS data. Results for k-means were also applied to kernel k-means. The current comprehension of k-means proceeds. Finally, numerical simulation studies are given, and the performance of the k-means for high-dimensional data is discussed.

E0588: **Kick-one-out-based variable selection method for Euclidean distance-based classifier in high-dimensional settings**
Presenter: **Tomoyuki Nakagawa**, Meisei University, Japan
Co-authors: Hiroki Watanabe, Masashi Hyodo

Classification is considered by using Euclidean distance-based classifier in high-dimensional data. The most important and difficult part of discriminant model building is selecting the appropriate variables for discriminant analysis. If the non-redundant variables that affect the classification rule are omitted, the expected probability of misclassification (EPMC) is large. On the other hand, EPMC may be large even if the redundant variables, which do not affect the classification rule, are included. Kick-One-Out(KOO)-based criteria are established for selecting non-redundant variables for the Euclidean distance-based classifier, and the consistency of the proposed variable selection in high dimensional settings is proved. The theoretical results are derived without assuming homogeneity of covariance matrices and multivariate normality for the group-conditional distribution.

E0814: **Sure screening for interaction effect in generalized linear models**
Presenter: **Yuta Umezu**, Nagasaki University, Japan

In regression models, detecting interactions between several covariates is an important task in many areas. However, when interaction effects are modelled, the model's size becomes vastly large even if the number of covariates is relatively small. Consequently, handling high dimensionality is needed, which means the number of parameters may be much larger than the sample size. Detecting such interactions in generalized linear models based on marginal screening are investigated. To this end, a marginal maximum likelihood estimator is considered in which the Lasso penalty is adopted only for the coefficient of the interaction term. As a result, a simple but effective screening rule can be obtained. Moreover, the screening rule enjoys good theoretical property, the so-called sure screening property; truly active interactions with probability converging to one under appropriate conditions can be detected. Several simulation studies and a real data example for checking the method's performance will also be presented.

E0707: **Linear hypothesis testing on mean vectors for factor model in high-dimensional settings**
Presenter: **Takahiro Nishiyama**, Senshu University, Japan
Co-authors: Masashi Hyodo

For high-dimensional data, a general linear hypothesis testing problem on mean vectors of several populations is discussed, which includes many existing hypotheses about mean vectors as special cases. For this problem, based on L^2-type statistic, a testing procedure that accommodates a low-dimensional latent factor model under heteroscedasticity is proposed for this problem. Under a high-dimensional asymptotic regime, combined with weak technical conditions, it is shown that null limiting distributions of the test statistics follow a weighted mixture of chi-square distributions. Also, an evaluation of the finite sample performance of the proposed tests by a simulation study is provided.

E0462: **Correlation matrix of factor model: Fluctuation of largest eigenvalue, scaling of bulk eigenvalues, and stock market**
Presenter: **Yohji Akama**, Tohoku University, Japan

Consider an N-dimensional sample of size T and a sample correlation matrix C. Suppose that N and T tend to infinity with T/N converging to a fixed finite constant $Q > 0$. If the population is a factor model, then the eigenvalue distribution of C almost surely converges weakly to Marcenko-Pastur distribution such that the index is Q and the scale parameter is the limiting ratio of the specific variance to the i-th variable in the limit o4f i. For an N-dimensional normal population with an equi-correlation coefficient r, which is a one-factor model, for the largest eigenvalue l of C, we prove that l/N converges to r almost surely. These results suggest an important role of an equi-correlated normal population and a factor model: the histogram of the eigenvalue of the sample correlation matrix of the returns of stock prices fits the density of Marcenko-Pastur distribution of index T/N and scale parameter $1 - l/N$. Moreover, the limiting distribution of the largest eigenvalue of a sample covariance matrix of an equi-correlated normal population is provided. The phase transition is discussed regarding the decay rate of the equi-correlation coefficient in N.

EO178 Room 709 NEW TOPICS IN MATHEMATICAL STATISTICS Chair: Yoichi Nishiyama

E0507: **The Dantzig selector for semiparametric models of stochastic processes**
Presenter: **Kou Fujimori**, Shinshu University, Japan
Co-authors: Koji Tsukuda

The sparse estimation problem is considered for models of stochastic processes with possibly infinite-dimensional nuisance parameters by using the Dantzig selector, which can be seen as an extension of the Z-estimator. When a consistent estimator for a nuisance parameter is obtained, it is possible to construct an asymptotically normal estimator for the parameter of interest under appropriate conditions. Motivated by this fact, the asymptotic behaviour of the Dantzig selector is established for models of ergodic stochastic processes with high-dimensional parameters of interest and possibly infinite-dimensional nuisance parameters. The applications for ergodic diffusion processes and ergodic time series, including integer-valued autoregressive models, are presented.

E0569: **A study on estimation in multivariate allometric regression**
Presenter: **Koji Tsukuda**, Kyushu University, Japan
Co-authors: Shun Matsuura

The multivariate allometric regression model is an extension of the allometric extension model to multivariate regression, which has been proposed previously. In the multivariate allometric regression model, it is assumed that the direction of the difference between the mean vectors of response variables for different values of explanatory variables always coincides with the first principal eigenvector of the covariance matrix of errors. Some estimators of the first principal eigenvector based on preliminary tests are proposed. The proposed estimators are also compared with conventional estimators.

E0709: **Evaluating the error probability of the spectral clustering algorithm in the allometric extension model**
Presenter: **Kohei Kawamoto**, Kyushu University, Japan
Co-authors: Yuichi Goto, Koji Tsukuda

The spectral clustering algorithm is often used as a binary clustering method for unclassified data by applying the principal component analysis. Several properties of the method have been studied under the assumption of the equality of two population covariance matrices. A non-asymptotic bound of the error probability of clustering is provided under the assumption of the allometric extension model; that is, the directions of the first eigenvectors of two covariance matrices and the direction of the difference of two mean vectors coincide.

E0712: **Estimation of n in the binomial (n, p) distribution with both parameters unknown**
Presenter: **Yoshiji Takagi**, Nara University of Education, Japan

Estimation of the population size n in the binomial (n, p) distribution with unknown success probability p has a long history of over eighty years, but easily computable and easily motivated estimators are still generally lacking. When p is apparently near zero, the familiar estimators obtained by frequentist methods are confronted by two difficulties, instability and underestimation. Indeed, the maximum likelihood estimator and the moment estimator can be extremely unstable in the sense that changing an observed success count s to $s+1$ can result in a massive change in the estimate of n. The sample maximum strongly underestimates the true n even for large sample sizes. The Bayesian approach is useful to overcome these difficulties, and several Bayesian estimators have been proposed. However, it is debatable how one should choose the parameters in the prior distribution. Here, the frequentist methods are reconsidered, and a new estimator constructed by four values is proposed: the sample minimum, the sample maximum, the sample mean and the sample variance. Last, some properties of this estimator are examined, including stability and less underestimation.

E0722: **Higher-order density derivative estimation for nonnegative data**
Presenter: **Yoshihide Kakizawa**, Hokkaido University, Japan

For the data supported on $[0, \infty)$, the so-called boundary bias problem is one of the interests, and asymmetric kernel density estimation has been well-studied. The asymmetric kernel method will be applied further to estimate higher-order density derivatives. Asymptotic bias and variance of the proposed higher-order density derivative estimator are derived, together with its M(I)SE property.

EC324 Room 201 APPLIED ECONOMETRICS I **Chair: Yongdai Kim**

E1021: **Exploring house price momentum in the US after the subprime mortgage crisis**
Presenter: **Heejoon Han**, Sungkyunkwan University, Korea, South
Co-authors: Pinshan Pan

The aim is to examine the relationship between house prices, rents, and user costs of housing in the United States from January 2009 to March 2022. First, the time-varying coefficient cointegration model is used to explain the long-run relationship and adopt an error correction model with endogenous regime switching, which turns out to fit the data better than existing models. The results show that the U.S. housing market has either a strong or weak house price momentum state after the subprime mortgage crisis. House price returns are more persistent in the strong momentum regime, and error correction is slower. The degree of house price momentum is estimated to be 1.104 and 0.339 in the strong and weak regimes, respectively. It is estimated that 74% of the data remains in the strong momentum regime. The extracted latent factor decides the regime of the housing market, and the adaptive lasso on the FRED-MD is run to find the link between the house price momentum and macroeconomic and financial variables.

E1096: **Bayesian multi-population Lee-Carter model applied to Japanese mortality data**
Presenter: **Hao Chen**, Hiroshima University, Japan
Co-authors: Haruhisa Nishino

The Lee-Carter model has been widely used for analyzing mortality. Bayesian approaches enable us to extend the model and achieve more flexible modelling. On the other hand, mortality research has recently focused on modelling data collected by categories such as regions. For analyzing these data, Bayesian multi-population mortality models are used, including a spatial conditional autoregressive (CAR) model, to prefecture-specific data from the Japanese Mortality Database (JMD). By comparing the models using the widely applicable information criterion (WAIC), the results show that the model containing the CAR outperformed other models without it. It suggests that the spatial effect should be included in analyzing prefecture-specific mortality data in Japan. The impact of great earthquakes on Japanese mortality rates is also examined, including the Great East Japan Earthquake and the Great Hanshin-Awaji Earthquake. The findings highlight the importance of considering the spatial effect in mortality research and demonstrate the benefits of using the multi-population mortality model with CAR parameters. The implications could help improve mortality forecasting and advance the understanding of mortality trends in Japan.

E1077: **An analysis of increased mortgage interest rates on the housing market in Germany using the BSTS Model**
Presenter: **Chong Dae Kim**, TH Koeln (Technische Hochschule Koeln), Germany

The short-term impact of the Ukrainian refugee crisis on the Polish rental housing market was discussed using the BSTS model. In Germany, the housing market benefited from low mortgage interest rates for a long period of time. The BSTS model is used because of the advantages already explained. The increase in the mortgage interest rates represents the intervention which can be located in spring 2022. Thus, the post-intervention period extends over three quarters. Since the effects of the increased mortgage interest rates on the housing market appear only after a certain period of time, the research question is whether these causal effects can be proven with the BSTS model over a limited period of time. The application of the BSTS model yields results on whether the impact of the increased mortgage interest rates on the housing market is statistically significant. Firstly, these results serve to discuss possible limitations of the model because of the limited time series. These limitations provide information on how soon, after an intervention, an analysis can be carried out using this model. Early conclusions about the causal effects of an intervention on the market allow for swift responses. Subsequently, if the causal effects of increased mortgage interest rates can actually be proven, the results also serve to validate this assumption by a mathematical model.

E0265: **Measuring contagion effects of crude oil prices on sectoral stock price indices in India**
Presenter: **Arvind Kumar Shrivastava**, Reserve Bank of India, India
Co-authors: Madhuchhanda Sahoo, Thangzason Sonna, Jessica Maria Anthony

The purpose is to explore the contagion effects of extreme changes in global crude oil prices on sectoral stock price indices in India. Using generalised Pareto distribution (GPD) for estimating excess returns or exceedances, i.e., deviations from thresholds, also a multinomial logit model (MNL) for assessing the probability of contemporaneous excess returns or co-exceedances, a significant likelihood of contemporaneous exceedances is found among ten sectoral stock price indices when faced with extreme changes in global crude oil prices pointing to the existence of a contagion effect. The evidence of positive co-exceedances is stronger, and the results are found more robust when relevant control variables are introduced, exchange rate returns (INR-USD), 10-year G-sec yield, and differential stock returns, (i.e., small firms minus big firms (SMB)). The contagion effect on all sectoral indices, irrespective of their direct and indirect exposure to oil price dynamics, highlights the need for hedging by investors as mere diversification of portfolios may not be sufficient to protect their assets from an adverse oil price shock.

E1248: **Analyzing market response to SFCR in European insurance with topic modeling and deep learning methods**
Presenter: **Chih-Chou Chiu**, National Taipei University of Technology, Taiwan
Co-authors: Ling-Jing Kao, Ting Kang Liu

A BERTopic-LSTM model was proposed to investigate the market response to Solvency and Financial Condition Reports (SFCRs) published by 27 publicly-traded insurance companies in the European Union in English between 2018 and 2021. The BERTopic method, a natural language processing (NLP) technique, was utilized to analyze the content of the SFCRs. A deep learning LSTM model was then developed to predict the changes in the stock yields of life insurance companies following the release of the SFCRs. The essential financial metrics and textual features of interest to investors were also examined. The results demonstrate that the proposed model surpasses competing machine learning models (MARS, Random Forest, and Gradient Boosting) in terms of prediction accuracy. Furthermore, the findings reveal that the corporate governance attributes of financial performance and risk management (topic 1) and operational management and resilience (topic 2) had the most significant influence on changes in stock yield prediction.

EC319 Room 203 MIXED MODELS **Chair: Pavlo Mozharovskyi**

E0216: **Asymptotic results for penalised quasi-likelihood estimation in generalised linear mixed models**
Presenter: **Xu Ning**, Australian National University, Australia
Co-authors: Francis Hui, Alan Welsh

Generalised linear mixed models (GLMMs) are widely used in the statistical analysis of clustered and correlated data. Other than in the special case of normal responses with an identity link, the likelihood of these models involves an intractable integral. One crucial and well-established method of avoiding this integral when fitting GLMMs is penalised quasi-likelihood (PQL) estimation. However, there are no formal asymptotic distribution results relating to PQL estimation for GLMMs in the literature. This gap is addressed by establishing large sample distributional results for PQL estimators of the parameters and random effects in independent-cluster GLMMs when the number of clusters and the cluster sizes go to infinity. This is done under two distinct frameworks: conditional on the random effects, treating them as fixed effects, and unconditionally, treating the random effects as random. Conditional on the random effects, it is shown that the PQL estimators are asymptotically normal around the true fixed and random effects. Unconditionally, the PQL estimator for fixed effects is proved that is asymptotically normal around the true fixed effects.

The asymptotic normality of the random effects estimator is proved, and the correct asymptotic distribution of the so-called prediction gap, which is not always normal, is derived. The finite sample performance of the theoretical results is verified through a simulation study.

E0355: **Increasing sample size asymptotic for two-way crossed mixed effect model**
Presenter: **Ziyang Lyu**, University of New South Wales, Australia

Asymptotic results for the maximum likelihood and restricted maximum likelihood (REML) estimators of the parameters in the two-way crossed mixed effect model for clustered data when the total of row, column and cell go to infinity. A set of mild conditions is given under which the estimators are shown to be asymptotically normal with an elegantly structured covariance matrix. There are no restrictions on the rate at which the cluster size tends to infinity, but it turns out to be essential to specify the regression function appropriately.

E1072: **Mixed effects models for large sized clustered extremes**
Presenter: **Koki Momoki**, Kagoshima University, Japan
Co-authors: Takuma Yoshida

Extreme value theory (EVT) provides an elegant mathematical tool for the statistical analysis of rare events. Typically, when data consists of multiple clusters, analysts want to preserve cluster information such as region, period, and group. To take into account the large-sized cluster information in extreme value analysis, the mixed effects model (MEM) is incorporated into the regression technique in EVT instead of the traditional approach, such as multivariate extreme value distribution. The MEM has been recognized not only as a model for clustered data but also as a tool for providing reliable estimates of large-sized clusters with small sample sizes. In EVT for rare event analysis, the effective sample size for each cluster is often small. Therefore, the MEM may also contribute to improving the predictive accuracy of extreme value analysis. However, to the best of our knowledge, the MEM has not yet been developed in the context of EVT. This motivates us to verify the effectiveness of the MEM in EVT through theoretical studies and numerical experiments, including application to real data for risk assessment of heavy rainfall in Japan.

E1090: **The R2D2 prior for generalized linear mixed models**
Presenter: **Eric Yanchenko**, North Carolina State University / Tokyo Institute of Technology, Japan

In Bayesian analysis, the selection of a prior distribution is typically done by considering each parameter in the model. While this can be convenient, it may be desirable to place a prior on a summary measure of the model instead of in many scenarios. A prior on the model fit is proposed, as measured by a Bayesian coefficient of determination (R^2), which then induces a prior on the individual parameters. This is achieved by placing a beta prior on R^2 and then deriving the induced prior on the global variance parameter for generalized linear mixed models. Closed-form expressions are derived in many scenarios and present several approximation strategies when an analytic form is not possible and/or to allow for easier computation. In these situations, approximating the prior is suggested by using a generalized beta prime distribution and a simple default prior construction scheme is provided. This approach is quite flexible and can be easily implemented in standard Bayesian software. Lastly, the method's performance is demonstrated on simulated data, where it particularly shines in high-dimensional examples and real-world data, which shows its ability to model spatial correlation in the random effects.

E1098: **MixTasteNet: A neural-embedded mixed logit model**
Presenter: **Alvaro Gutierrez Vargas**, KU Leuven, Belgium
Co-authors: Martina Vandebroek, Michel Meulders

The MixTasteNet model is proposed, a novel hybrid of an Artificial Neural Network (ANN) and a Mixed Logit (MIXL) model, for modelling "taste heterogeneity" in Discrete Choice Models (DCM). While conventional Multinomial Logit (MNL) models can incorporate observed heterogeneity by including interactions between alternative-specific regressors and individual characteristics, it is a cumbersome trial-and-error process that can rapidly increase the number of estimated parameters. In contrast, random coefficient models, such as MIXL models, aim to capture unobserved heterogeneity through distributional assumptions over the taste parameters. Hybrid models use individuals' characteristics to feed an ANN and produce heterogeneous taste parameters that are included in the utility specification. The MixTasteNet model goes a step further by simultaneously including random coefficients, which capture unobserved heterogeneity, and an ANN component, which captures the observed heterogeneity, in the utility specification. Notably, the proposed model is the first hybrid specification used in DCM that incorporates random coefficients and an ANN to model individuals' preferences. Finally, the MixTasteNet model accurately recovers the true models' parameters while achieving the predictability of the ground-truth model was demonstrated via simulation.

EC317 Room 708 FUNCTIONAL DATA ANALYSIS Chair: Alexander Petersen

E0459: **Additive regression with general imperfect variables**
Presenter: **Jeong Min Jeon**, Seoul National University, Korea, South
Co-authors: Germain Van Bever

An additive model is introduced, where the response variable is Hilbert-space-valued, and predictors are multivariate Euclidean, and both are possibly imperfectly observed. Considering Hilbert-space-valued responses allows us to cover Euclidean, compositional, functional and density-valued variables. By treating imperfect responses, functional variables taking values in a Riemannian manifold and the case where only a random sample can be covered from a density-valued response is available. Dealing with imperfect predictors allows us to cover various principal component and singular component scores obtained from Hilbert-space-valued variables. The smooth back-fitting method is used to estimate the additive model having such variables. Asymptotic properties of the regression estimator are provided, and a numerical study is presented.

E0531: **Riemannian functional regression and reproducing kernel tensor Hilbert spaces**
Presenter: **Ke Yu**, University of Oxford, United Kingdom
Co-authors: James Taylor

In many scientific fields, data arise in the form of smooth functions on Riemannian manifolds. Analyzing the relationship between a Riemannian functional response and a Riemannian functional predictor has become increasingly important. A Riemannian function-on-function regression model under the reproducing kernel tensor Hilbert space (RKTHS) framework is developed. As an extension of vector-valued reproducing kernel Hilbert spaces, the RKTHS we construct consists of functions taking values in tangent spaces along a curve on a manifold and is able to capture the intrinsic geometry of the manifold. The estimator of the regression coefficient achieves the optimal rate of convergence in mean prediction is proven. Moreover, a method is proposed to compare objects from different tensor Hilbert spaces based on Hilbert manifolds. Potential problems caused by nonnegative sectional curvatures of manifolds are also studied. Simulation studies demonstrate the numerical advantages of the RKTHS-based approach over the function-on-function regression based on Riemannian functional PCA. The proposed method is applied to tropical cyclone data to predict trajectories and brain imaging data of preterm and full-term infants in the Developing Human Connectome Project (dHCP) to study the linear relationship between homologous white matter fibre tracts in two hemispheres of the brain.

E0615: **Functional principal component analysis for partially observed elliptical process**

Presenter: **Yaeji Lim**, Chung-Ang University, Korea, South

The robust estimators of principal components are presented for partially observed functional data with heavy-tail behaviours, where sample trajectories are collected over individual-specific subinterval(s). The partially sampled trajectories are considered to be the filtered elliptical process by the missing indicator process, and implementing the robust functional principal component analysis under this framework is proposed. The proposed method is computationally efficient and straightforward by estimating the robust correlation function based on the pairwise covariance computation combined with M-estimation. The asymptotic consistency of the estimators is established under general conditions. The superior performance of our method in approximating the subspace of the data and reconstruction of full trajectories is demonstrated in simulation studies. Then the proposed method is applied to hourly monitored air pollutant data containing anomaly trajectories with random missing segments.

E1159: **Simultaneous clustering and dimensionality reduction of functional data**

Presenter: **Roberto Rocci**, Sapienza University of Rome, Italy

Co-authors: Stefano Antonio Gattone

A new technique for simultaneous clustering and dimensionality reduction of functional data is proposed. The observations are projected into a low-dimensional subspace and clustered by means of functional K-means. The subspace and the partition are estimated simultaneously by minimizing the within deviance in the reduced space. This allows us to find new dimensions with a very low within deviance, which should correspond to a high level of discriminant power. However, in some cases, the total deviance explained by the new dimensions is so low as to make the subspace and, therefore, the partition identified in it insignificant. In order to overcome this drawback, a penalty equal to the negative total deviance in the reduced space is added to the loss. In this way, subspaces with low deviance are avoided. It is shown how several existing methods are particular cases of the proposal simply by varying the weight of the penalty. A further penalty is added in order to take into account the functional nature of the data by smoothing the centroids, and an alternating least squares algorithm is introduced to compute model parameter estimates. An application to real and simulated data shows the effectiveness of the proposal.

E1239: **Functional adaptive double-sparsity estimator for functional linear regression with multiple functional covariates**

Presenter: **Xinyue Li**, City University of Hong Kong, Hong Kong

Wearable sensors have been increasingly used in health monitoring and early anomaly detection. Wearable devices can collect objective and continuous information on physical activity and vital signs and have great potential in studying the association with health outcomes. However, how to effectively analyze high-frequency multi-dimensional sensor data is challenging. A new Functional Adaptive Double-Sparsity Estimator (FadDoS) based on functional regularization of sparse group lasso with multiple functional predictors is proposed, which can achieve global sparsity via functional variable selection and local sparsity via zero-subinterval identification within coefficient functions. It is proved that the FadDoS estimator converges at a bounded rate and satisfies the oracle property under mild conditions. Extensive simulation studies confirm the theoretical properties and exhibit excellent performances compared to existing approaches. Application to a Kinect sensor study that utilized an advanced motion sensing device tracking human multiple joint movements and conducted among community-dwelling elderly demonstrates how FadDoS can effectively characterize the detailed association between joint movements and physical health assessments. The proposed method is not only effective in Kinect sensor analysis but also applicable to broader fields, where multi-dimensional sensor signals are collected simultaneously, to expand the use of sensor devices in health studies.

Wednesday 02.08.2023 08:15 - 10:20 Parallel Session F – EcoSta2023

EO252 Room 03 RECENT ADVANCES IN STATISTICAL THEORY AND METHODS Chair: Arlene Kyoung Hee Kim

E0359: **Unpaired regression for a discrete response via Poisson quantiles matching**
Presenter: **Hyungjun Lim**, Korea University, Korea, South
Co-authors: Arlene Kyoung Hee Kim

Analyzing the data collected from different sources requires unpaired data analysis to account for the absence of correspondence between the response variable Y and explanatory variables X. Several attempts have been made to analyze continuous Y, but the response variable of interest may follow a discrete distribution, which previous methodologies have overlooked. To address these limitations, Poisson quantile matching estimation (PQME) is proposed, the first unpaired data analysis method designed to examine the discrete response variable Y and the unpaired continuous explanatory variable X. Using their order statistics, the PQME method matches the linear combination of explanatory variables to $\ln(Y)$. An effective algorithm and simulation results are presented, along with the convergence results. The practical application of PQME is illustrated by locating the ideal site for a new facility using real data.

E0767: **A comprehensive framework for investigating multiple latent class variables**
Presenter: **Youngsun Kim**, Korea University, Korea, South
Co-authors: Hwan Chung

Latent class analysis (LCA) is a popular method for population segmentation, but it may not be sufficient for capturing complex population structures that require multiple latent class variables. Several approaches, such as Latent Transition Analysis (LTA), Latent Class Profile Analysis (LCPA), Joint Class Analysis (JLCA), and Joint Latent Class Profile Analysis (JLCPA), have been developed to explore the association among multiple latent class variables. A new framework is proposed, called the Structural Latent Class Model (SLCM), which integrates these existing LCA variants into a single framework by linking multiple latent class variables via transition matrices. A user-friendly R package for implementing the models has also been developed.

E0522: **Variable selection for AUC-optimizing classification in diverging dimensions**
Presenter: **Hyungwoo Kim**, Pukyong National University, Korea, South

The purpose is to investigate the asymptotic behaviours of the estimator of the AUC-optimizing classification penalized by the smoothly clipped absolute deviation (SCAD) penalty. First, the AUC consistency over the linear function class is studied. Then it is proven that the SCAD-penalized estimator possesses the oracle property under both cases where the predictor dimension is fixed and diverges to infinity. For choosing the regularization parameter in SCAD-penalized AUC-optimizing classification, a BIC-type information criterion is proposed,, shown to capture the true model consistently. The technical proofs are based on the theory of U-processes. In addition, an applicable computation algorithm has been developed to estimate the SCAD-penalized estimator. Both simulated and real data analysis results demonstrate the promising performance of the proposed method in terms of variable selection and prediction.

E0334: **Nonparametric reduced-rank estimation of multiple regression functions**
Presenter: **Kwan-Young Bak**, Sungshin Women's University, Korea, South
Co-authors: Ja Yong Koo

A multi-task nonparametric regression problem in which the underlying functions possess a low-rank structure is examined. A nonparametric function estimation method based on the nuclear norm penalization (NNP) approach is proposed to incorporate the low-dimensional structure in the recovery of multiple functions. This leads to a nonparametric version of the reduced rank regression estimator under the multi-task learning framework. Numerical studies are provided to illustrate the efficiency of the proposed method. The results show that the information pooling across multiple experiments based on the low-rank structure can significantly outperform the separate estimation method. Regarding the theoretical aspect, a non-asymptotic oracle-type inequality is first obtained to study the properties of the reduced rank estimator. Using the inequality, the proposed method's minimaxity and rank identification consistency is proved.

E0696: **Interval-censored linear quantile regression**
Presenter: **Taehwa Choi**, Duke University, United States
Co-authors: Seohyeon Park, Hunyong Cho, Sangbum Choi

Censored quantile regression has emerged as a prominent alternative to classical Cox's proportional hazards model or accelerated failure time model in both theoretical and applied statistics, as it enables researchers to investigate the complete distribution of survival responses with respect to a set of covariates. While quantile regression has been extensively studied for right-censored survival data, the survival analysis literature methodologies for analyzing interval-censored data remain limited. A novel local weighting approach is proposed for estimating linear censored quantile regression with various types of interval-censored survival data. The regression parameter's estimation equation and the corresponding convex objective function can be constructed as a weighted average of quantile loss contributions at two interval endpoints. The weighting components are nonparametrically estimated using local kernel smoothing or ensemble machine-learning techniques. A modified EM algorithm for nonparametric distribution mass for interval-censored data is employed by introducing subject-specific latent Poisson variables to estimate the nonparametric maximum likelihood estimation. The proposed method's empirical performance is demonstrated through extensive simulation studies and real data analyses of two HIV/AIDS datasets.

EO008 Room 04 RECENT ADVANCES IN CAUSAL INFERENCES Chair: Zheng Zhang

E0172: **A conditional linear combination test with many weak instruments**
Presenter: **Yichong Zhang**, Singapore Management University, Singapore

A linear combination of jackknife Anderson-Rubin, jackknife Lagrangian multiplier (LM), and orthogonalized jackknife LM tests for inference in IV regressions are considered with many weak instruments and heteroskedasticity. The weights in the linear combination are chosen based on a decision-theoretic rule that is adaptive to the identification strength. Under both weak and strong identifications, the proposed test controls the asymptotic size and is admissible among specific classes of tests. Under strong identification, the linear combination test has optimal power against local alternatives. Simulations and an empirical application to Angrist and Krueger in the 1991 dataset confirm the good power properties of the test.

E0224: **The synthetic instrument**
Presenter: **Linbo Wang**, University of Toronto, Canada
Co-authors: Dingke Tang, Dehan Kong, Linbo Wang

In many observational studies, researchers are interested in studying the effects of multiple exposures on the same outcome. Unmeasured confounding is a key challenge in these studies as it may bias the causal effect estimate. To mitigate the confounding bias, a novel device called the synthetic instrument is introduced to leverage the information contained in multiple exposures for causal effect identification and estimation. It is shown that under linear structural equation models, the problem of causal effect estimation can be formulated as an ℓ_0 penalization problem, and hence can be solved efficiently using off-the-shelf software. Simulations show that our approach outperforms state-of-the-art methods in low- and high-dimensional settings. The study further illustrates our method using a mouse obesity dataset.

E1273: **Causal inference with invalid instruments: Exploring nonlinear treatment models with machine learning**
Presenter: **Zijian Guo**, Rutgers University, United States

Causal inference for observational studies with possibly invalid instrumental variables is discussed. A novel methodology called two-stage curvature identification (TSCI) is proposed, which explores the nonlinear treatment model with machine learning and adjusts for different forms of violating the instrumental variable assumptions. The success of TSCI requires the instrumental variable's effect on treatment to differ from its violation form. A novel bias correction step is implemented to remove bias resulting from the potentially high complexity of machine learning. The proposed TSCI estimator is shown to be asymptotically unbiased and normal even if the machine learning algorithm does not consistently estimate the treatment model. A data-dependent method is designed to choose the best among several candidate violation forms. TSCI is applied to study the effect of education on earnings.

E1293: **Nonparametric estimation of general heterogeneous causal effects with covariate measurement error**
Presenter: **Wei Huang**, University of Melbourne, Australia
Co-authors: Zheng Zhang, Haoze Hou

Estimation of heterogeneous treatment effect plays a central role in the area of economics, social science and psychology, among others. Existing literature focuses on studying the conditional average treatment effect (CATE), assuming all variables are measured without errors. A unified framework is proposed for estimating general heterogeneous treatment effects (GHTE) when the conditioning covariates are exposed to classical measurement errors. The framework includes conditional average, quantile and asymmetric least squares treatment effects as special cases. Under the unconfoundedness condition, the pointwise asymptotic distribution is provided with influence functions and uniform confidence bands, which are useful for statistical inference in practice.

EO139 Room Virtual R01 RECENT ADVANCES IN THE ANALYSIS OF DATA WITH COMPLEX STRUCTURES Chair: Yuhang Xu

E0239: **On spatial generalized autoregressive conditional heteroskedasticity varying coefficient models**
Presenter: **Jingru Mu**, Kansas State University, United States
Co-authors: Liying Jin

A new volatility model is proposed by allowing spatially varying coefficients in spatial generalized autoregressive conditional heteroskedasticity (SGARCH) models. This model captures volatility behaviours over space and investigates the relationship between some explanatory variables and the volatility at each location. A two-stage quasi-likelihood maximization via BPST is developed to estimate the model over a complicated domain. The theoretical properties of the proposed estimators are also presented. Both simulation studies and real-data applications are conducted to demonstrate the performance of our approach.

E0241: **Shifting-corrected regularized regression for 1H NMR metabolomics identification and quantification**
Presenter: **Yuhang Xu**, Bowling Green State University, United States
Co-authors: Thao Vu, Yumou Qiu

The process of identifying and quantifying metabolites in complex mixtures plays a critical role in metabolomics studies to obtain an informative interpretation of underlying biological processes. Manual approaches are time-consuming and heavily reliant on the knowledge and assessment of nuclear magnetic resonance (NMR) experts. A shifting-corrected regularized regression method is proposed, automatically identifying and quantifying metabolites in a mixture. A detailed algorithm is also proposed to implement the proposed method. Using a novel weight function, the proposed method can detect and correct peak shifting errors caused by fluctuations in experimental procedures. Simulation studies show that the proposed method performs better in identifying and quantifying metabolites in a complex mixture. This approach uses experimental and biological NMR mixtures to demonstrate real data applications.

E0242: **Clustered coefficient regression models for Poisson process with an application to seasonal warranty claim data**
Presenter: **Xin Wang**, San Diego State University, United States
Co-authors: Xin Zhang, Zhengyuan Zhu

Motivated by a product warranty claims data set, clustered coefficient regression models are proposed in a non-homogeneous Poisson process for recurrent event data. The proposed method referred to as CLUPP, can simultaneously estimate the group structure and parameters. The proposed method uses a penalized regression approach to identify the group structure. Numerical studies show that the proposed approach can identify the group structure well and outperforms traditional methods such as hierarchical clustering and K-means. Theoretical properties are also established, which show that the proposed estimators can converge to true parameters in high probability. The proposed methods are ultimately applied to the product warranty claims data set, achieving better prediction than the state-of-the-art methods.

E0387: **Comparing methods for determining power priors based on different congruence measures**
Presenter: **Jing Zhang**, Miami University, United States
Co-authors: John Bailer, Ainsley Helling

A Ceriodaphnia dubia (C. dubia) reproduction test is often used to evaluate the sublethal toxicity of water effluents or chemicals. In a C. dubia reproduction test, organisms are exposed to varying concentration levels of the toxicant or other adverse treatment and the number of young after a given experiment period is recorded. The reproduction response is modelled as a function of the concentration, and the concentration associated with specified levels of estimated adverse effect is used in risk management to analyse the experiment outcomes. While aquatic toxicity analyses often focus on outcomes from the current experiment, laboratories commonly have a history of conducting such experiments using the same species, following a similar experimental protocol. So it is often reasonable to believe that the same underlying biological process generates the historical and current experiments. In the present study, using a calibrated power prior approach is proposed to incorporate historical control outcomes as prior input and compare the behaviour of the method when different congruence measures are used to determine the amount of historical input that will be incorporated. Simulation results show that three of the congruence measures show attractive features in practice. Using the calibrated power priors would improve precision and the bias of the potency estimates.

E0420: **Longitudinal disparity decomposition under the varying-coefficient framework**
Presenter: **Seonjin Kim**, Miami University, United States
Co-authors: Sang Kyu Lee, Mi-Ok Kim, Hyokyoung Hong

A varying-coefficient longitudinal disparity model is proposed for analyzing disparity between population groups. The coefficients in the longitudinal model vary over time and modifier. A sophisticated decomposition with respect is presented to the modifier based on the proposed model. The new decomposition shows not only the direct effect of the modifier but also the indirect effect of the modifier through other covariates on the response variable. In addition, it enables us to investigate the traditionally explained disparity by group differences and unexplained disparity by the covariates. The approach is applied to assess a longitudinal disparity of fetal growth.

EO090 Room Virtual R02 ESTIMATION OF EIGENVECTORS AND COVARIANCE MATRICES IN HIGH DIMENSIONS Chair: Lisa Goldberg

E0304: **James-Stein Estimator of moderately-spiked leading eigenvector**
Presenter: **Sungkyu Jung**, Seoul National University, Korea, South

Recently, a James-Stein shrinkage (JS) estimator has gained attention as a powerful tool for estimating the -leading eigenvector of covariance matrices. The efficacy of the JS estimator has been demonstrated under a strongly-spiked leading eigenvalue model, using the high-dimensional, low-sample-size (HDLSS) asymptotic regime, where the number of variables increases while the sample size remains fixed. We extend the application of the JS shrinkage to the regime of moderately-spiked leading eigenvalues and reveal a key condition involving a signal-to-noise ratio, for the JS estimator to be useful. Furthermore, we develop shrinkage estimators for principal component variance and scores, enabling their application in high-dimensional principal component analysis.

E0758: **Issues in large covariance matrix estimation for portfolio risk prediction**
Presenter: **Stjepan Begusic**, Unversity of Zagreb, Croatia

Most studies considering the problem of estimating large covariance matrices of asset returns from a short time window (high dimension low sample size (HDLSS) regime) focus on portfolio optimization applications. It is known that mean-variance optimization acts as error maximization (especially in the HDLSS regime), which has been mitigated by certain estimators, such as those based on the spiked covariance model. By imposing factor models or shrinking the covariance estimates towards them, the robustness of optimized portfolios can be improved, and their out-of-sample risk reduced. However, correcting the optimization error might lead to introducing errors in other applications, most notably those in portfolio risk prediction. The aim is to focus on these applications, specifically the estimation of portfolio variance in a spiked covariance model and the issues which arise with different covariance matrices in the HDLSS regime. The properties of the risk prediction errors and the effects of different eigenstructure shrinkage methods on these errors are considered. An experimental study is presented, for a range of dimensionality scenarios and various portfolios, together with some insights for practitioners and directions for future work.

E1227: **James-Stein for eigenvectors with applications to constrained optimization**
Presenter: **Lisa Goldberg**, University of California, Berkeley, United States
Co-authors: Alec Kercheval, Hubeyb Gurdogan

A recipe for estimating covariance matrices tailored to constrained optimization problems is provided, resulting in optimized portfolios with low variance. The recipe relies on recent research that identifies and corrects bias, such as excess dispersion, in the leading sample eigenvector of a factor-based covariance matrix estimated from a high-dimension low sample size (HL) data set. It is shown that eigenvector bias can have a substantial impact on variance-minimizing optimization in the HL regime, while bias in estimated eigenvalues may have little effect. The estimated covariance matrix is obtained with a data-driven eigenvector shrinkage operator called JamesStein for eigenvectors (JSE), which corrects bias that is selectively distorted by constrained optimization. The shrinkage operator is named to emphasize its profound parallels to classical JamesStein (JS) shrinkage operators for a collection of averages.

E1232: **Optimal regularization of the first principal component**
Presenter: **Youhong Lee**, University of California, Santa Barbara, United States

The concept of regularization, which combines a simple structured target with traditional estimators, is widely used in high-dimensional data analysis. A novel regularization technique and its efficient machine learning algorithm, termed direction-regularized principal component analysis (drPCA), are introduced. This method addresses the PCA problem, aiming to identify the direction of maximum variance in the data while adhering to a predefined target direction. Using the high-dimensional, low-sample size framework, an asymptotic analysis of the solution is performed, which results in an optimal tuning parameter that minimizes an asymptotic loss function. The data rapidly acquires the estimator corresponding to the optimal tuning parameter. Furthermore, it is demonstrated that under certain covariance structures, the estimator is equivalent to both the Ledoit-Wolf constant correlation shrinkage estimator and a recently proposed James-Stein estimator for the first principal component.

E1257: **Beyond James-Stein estimation for PCA**
Presenter: **Alexander Shkolnik**, University of California, Santa Barbara, United States

Recent progress on James-Stein estimation of principal components and, more generally, of singular vectors of random data matrices is surveyed. Connections to Bayesian methods, regularization, and Ledoit-Wolf covariance estimation are discussed. Several results on the convergence properties of the James-Stein estimator for the leading singular vector pair are presented. The topic of the inadmissibility of Principal Component Analysis (PCA), as well as the inadmissibility of the James-Stein estimator for PCA, are also discussed.

EO052 Room 102 RECENT DEVELOPMENTS IN ESTIMATION METHODS: THEORY AND APPLICATIONS Chair: Zhihua Su

E1011: **Comparing baseball players across eras via the novel full house model**
Presenter: **Daniel Eck**, University of Illinois, United States

A new methodological framework for era-adjusting baseball statistics is motivated. The proposed methodology is a crystallization of the conceptual ideas put forward by Stephen Jay Gould. This methodology is named the Full House Model in his honour. The Full House Model works by balancing the achievements of Major League Baseball (MLB) players within a given season and the size of the MLB-eligible population. The utility of the Full House Model in an application of comparing baseball players' performance statistics across eras is demonstrated. The results reveal a radical reranking of baseball's greatest players that is consistent with what one would expect under a sensible uniform talent generation assumption. Most importantly, it is found that the greatest modern players, including several African American, Latino, and Asian players, now sit atop the greatest all-time lists of historical baseball players, while conventional wisdom ranks such players lower. The conclusions largely refute a consensus of baseball greatness that is reinforced by nostalgic bias, recorded statistics, and statistical methodologies, which it is argued are not suited to the task of comparing players across eras.

E1015: **Penalized synthetic control method for truncated data with latent clustering**
Presenter: **Bikram Karmakar**, University of Florida, United States
Co-authors: Gourab Mukherjee, Wreetabrata Kar

A novel penalized synthetic control (SC) method is developed that accommodates latent structures often inherent in panel data structures. Using the proposed method, the effects of the passage of a medical marijuana law (MML) are studied by a state on direct payments to physicians. As an example of this latent structure, consider a physician who receives payments periodically, every six months, while another physician receives payments every five months. Direct use of an SC method while comparing these two physicians ignores these distinct latent patterns and thus will result in interpolation bias in the estimated effect. Under a truncated flexible additive mixture model, it is theoretically established that the SC method has uncontrolled maximal risk without a penalty; by contrast, the proposed penalized method provides efficient estimates. The analysis also estimates heterogeneous causal effects. Using the proposed method, a significant decrease in direct payments from opioid manufacturers is found to pain medicine physicians as an effect of MML passage. Evidence is provided that this decrease is due to the availability of medical marijuana as a substitute. Finally, the substitution effect is comparatively higher for female physicians and in localities with higher white, less affluent, and more working-age populations.

E1031: **Nonconvex-regularized integrative sufficient dimension reduction for multi-source data**
Presenter: **Wei Qian**, University of Delaware, United States
Co-authors: Shanshan Ding

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.

E1044: **Estimating heterogeneous causal mediation effects with bayesian decision tree ensembles**
Presenter: **Antonio Linero**, University of Texas at Austin, United States
Co-authors: Angela Ting

The causal inference literature has increasingly recognized that explicitly targeting treatment effect heterogeneity can lead to improved scientific understanding and policy recommendations. Towards the same end, studying the causal pathway connecting the treatment to the outcome can also be useful. These problems are addressed in the context of causal mediation analysis. A varying coefficient model based on Bayesian additive regression trees is introduced to identify and regularize heterogeneous causal mediation effects; analogously with linear structural equation models, these effects correspond to covariate-dependent products of coefficients. It is shown that, even on large datasets with few covariates, LSEMs can produce highly unstable estimates of the conditional average direct and indirect effects, while the Bayesian causal mediation forests model produces stable estimates. It is found that the approach is conservative, with effect estimates "shrunk towards homogeneity". The method's salient properties are examined using data from the Medical Expenditure Panel Survey and empirically grounded simulated data. Finally, the purpose is to show how the model can be combined with posterior summarization strategies to identify interesting subgroups and interpret the model fit.

E1286: **Response variable selection in multivariate linear regression**
Presenter: **Zhihua Su**, University of Florida, United States
Co-authors: Kshitij Khare

Response variable selection and subsequent estimation of the regression coefficients in multivariate linear regression are discussed. Because of the asymmetric roles of the predictors and responses in regression, response variable selection is markedly different from the usual predictor variable selection. When a response is inferred to have coefficients zero, it should not be simply removed from subsequent estimation. Instead, its relationship is analyzed with the responses that have nonzero coefficients, called dynamic responses. If it is correlated with the dynamic responses given all other responses, it should be retained to improve the estimation efficiency of the nonzero coefficients as an ancillary statistic. Otherwise, it can be removed from further inference (leading to significant resource savings in high-dimensional settings), called a static response. Therefore, the responses can be classified into three categories: dynamic, ancillary, and static. An algorithm is derived to identify these response variables and provide an estimator of the regression coefficients based on the selection result.

EO050 Room 201 WEIGHTING AND DYNAMIC APPROACHES TO CAUSAL INFERENCE Chair: Luke Keele

E0270: **Balanced and robust randomized treatment assignments: The finite selection model**
Presenter: **Jose Zubizarreta**, Harvard University, United States
Co-authors: Ambarish Chattopadhyay, Carl Morris

The Finite Selection Model (FSM) was developed previously for the design of the RAND Health Insurance Experiment (HIE), one of the largest and most comprehensive social science experiments conducted in the U.S. In the FSM, a treatment group at each of its turns selects the available unit that maximally improves the combined quality of its resulting group of units according to a common optimality criterion. In the HIE and beyond, the FSM is revisited, formalized, and extended as a general tool for experimental design. Leveraging the idea of D-optimality, a new selection criterion in the FSM is proposed and analyzed. The FSM using the D-optimal selection function has no tuning parameters, is affine invariant, and can retrieve classical designs such as randomized block and matched-pair designs. For multi-arm experiments, algorithms are proposed to generate a selection order of treatments. FSM's performance is demonstrated in a case study based on the HIE, a simulation study, and in ten randomized studies from the health and social sciences. On average, the FSM achieves 68% and 56% better covariate balance than complete randomization and rerandomization in a typical study. The FSM is recommended to be considered in experimental design for its conceptual simplicity, efficiency, and robustness.

E0274: **Approximate balancing weights for clustered observational study designs**
Presenter: **Luke Keele**, University of Pennsylvania, United States

In a clustered observational study, treatment is assigned to groups, and all units within the group are exposed to the treatment. A new method is developed for statistical adjustment in clustered observational studies using approximate balancing weights, a generalization of inverse propensity score weights that solve a convex optimization problem to find a set of weights that directly minimize a measure of covariate imbalance, subject to an additional penalty on the variance of the weights. The approximate balancing weights optimization problem is tailored to both adjustment sets by deriving an upper bound on the mean square error for each case and finding weights that minimize this upper bound, linking the level of covariate balance to a bound on the bias. The procedure is implemented by specializing the bound to a random cluster-level effects model. This leads to a variance penalty that incorporates the signal signal-to-noise ratio and penalizes the weight on individuals and the total weight on groups differently according to the intra-class correlation.

E0593: **Bayesian nonparametric methods for longitudinal mediation with informative continuous-time treatment decisions**
Presenter: **Jason Roy**, Rutgers University, United States
Co-authors: Arman Oganisian

The time between treatment courses might be informative in many clinical settings, such as cancer chemotherapy. In such settings, when there are questions about treatment's direct and indirect effects, available statistical methods are limited. Flexible Bayesian models are proposed for the continuous time decision process, time-varying mediator, and time-varying covariates. Then a g-computation approach is used to obtain the posterior distribution for the direct and indirect effects. The motivating example involves quantifying the contribution of chemotherapy-associated sepsis (time-varying mediator) to transient cardiac toxicity, which may help mitigate premature discontinuation of anthracycline chemotherapy agents. The performance of the models is assessed via simulations, and it is applied to data from a study of acute myeloid leukaemia.

E0599: **Bayesian semiparametric models for informatively timed, dynamic treatments with incomplete covariate trajectories**
Presenter: **Arman Oganisian**, Brown University, United States
Co-authors: Jason Roy

A Bayesian semiparametric model is developed for assessing the impact of dynamic treatment rules (DTRs) on survival. The motivating data are from a phase III clinical trial in which patients diagnosed with pediatric acute myeloid leukaemia (AML) move through a sequence of four treatment courses. At each course, a decision is made to administer anthracyclines (ACT). Since ACT is cardiotoxic, left ventricular ejection fraction (EF) is sometimes - but not always - measured beforehand to help inform the ACT decision. Inconsistent assessment leads to incomplete information on EF trajectories, a key tailoring variable. Moreover, patients 1) initiate each course at different times depending on the speed of recovery from previous courses, 2) may die or 3) be withdrawn from the study before ever completing the full sequence. The problem is framed in terms of a joint treatment-monitoring DTR that outputs both an EF monitoring decision and an ACT treatment decision. Gamma Process priors are used to flexibly model continuous-time transitions between treatment courses and death under hypothetical DTRs. A g-computation procedure simulates the transition process under hypothetical DTRs and computes posterior marginal survival probabilities.

E0894: **Longitudinal causal analysis of HIV antiretroviral therapy effects on weight gain**
Presenter: **Andrew Spieker**, Vanderbilt University Medical Center, United States

Prior studies have examined the effects of various antiretroviral therapy regimens and their impacts on subsequent weight gain in persons living with HIV cross-sectionally. However, large observational cohorts of persons living with HIV often feature antiretroviral treatment trajectories that are unstable over time. Traditional approaches to analyzing longitudinal causal effects can sometimes be criticized for over-reliance on stringent parametric assumptions. A flexible semi-parametric approach based on cumulative probability models examines the longitudinal effects of core and ancillary agents comprising antiretroviral therapies on weight gain in a large cohort of persons living with HIV. More robust evidence supports the hypothesis that modern integrase strand transfer inhibitors and tenofovir alafenamide are associated with greater mean weight gain than other core and ancillary antiretroviral agents.

EO169 Room 203 CAUSAL MACHINE LEARNING WITH HIGH DIMENSIONAL MODELING Chair: Tiffany Tang

E0682: **Limit theorems for semidiscrete optimal transport maps**
Presenter: **Kengo Kato**, Cornell University, United States
Co-authors: Ziv Goldfeld, Ritwik Sadhu

Statistical inference is studied for the optimal transport (OT) map (also known as the Brenier map) from a known absolutely continuous reference distribution onto an unknown finitely discrete target distribution. Limit distributions are derived for the L^p-estimation error with arbitrary $p \in [1,\infty)$ and for linear functionals of the empirical OT map. The former has a non-Gaussian limit, while the latter attains asymptotic normality. For both cases, consistency of the nonparametric bootstrap is also established. The derivation of the limit theorems relies on new stability estimates of functionals of the OT map with respect to the dual potential vector, which could be of independent interest.

E0498: **Optimal nonparametric inference with two-scale distributional nearest neighbors**
Presenter: **Lan Gao**, University of Tennessee Knoxville, United States
Co-authors: Yingying Fan, Jinchi Lv, Emre Demirkaya, Patrick Vossler, Jingbo Wang

The weighted nearest neighbours (WNN) estimator has been popularly used as a flexible and easy-to-implement nonparametric tool for mean regression estimation. The bagging technique is an elegant way to form WNN estimators with weights automatically generated to the nearest neighbours; the resulting estimator as the distributional nearest neighbours (DNN) for easy reference is named. Yet, there is a lack of distributional results for such an estimator, limiting its application to statistical inference. Moreover, when the mean regression function has higher-order smoothness, DNN does not achieve the optimal nonparametric convergence rate, mainly because of the bias issue. An in-depth technical analysis of the DNN is provided, based on which a bias reduction approach for the DNN estimator is suggested by linearly combining two DNN estimators with different subsampling scales, resulting in the novel two-scale DNN (TDNN) estimator. The two-scale DNN estimator is proven to enjoy the optimal nonparametric convergence rate in estimating the regression function under the fourth-order smoothness condition. Further beyond estimation, it was established that the DNN and two-scale DNN are asymptotically normal as the subsampling scales and sample size diverge to infinity. The theoretical results and appealing finite-sample performance of the suggested two-scale DNN method are illustrated with simulation examples and a real data application.

E0566: **GhostKnockoff inference empowers identification of putative causal variants in genome-wide association studies**
Presenter: **Zihuai He**, Stanford University, United States

Recent advances in genome sequencing and imputation technologies provide an exciting opportunity to study the contribution of genetic variants to complex phenotypes comprehensively. However, our ability to translate genetic discoveries into mechanistic insights remains limited at this point. An efficient knockoff-based method, GhostKnockoff, for genome-wide association studies (GWAS) leads to improved power and ability to prioritize putative causal variants relative to conventional GWAS approaches. The method requires only Z-scores from conventional GWAS and hence can be easily applied to enhance existing and future studies. The method can also be applied to a meta-analysis of multiple GWAS, allowing for arbitrary sample overlap. Its performance is demonstrated using empirical simulations and two applications: (1) a meta-analysis for Alzheimer's disease comprising nine overlapping large-scale GWAS, whole-exome and whole-genome sequencing studies and (2) an analysis of 1403 binary phenotypes from the UK Biobank data in 408,961 samples of European ancestry. Our results demonstrate that GhostKnockoff can identify putatively functional variants with weaker statistical effects that conventional association tests miss.

E0604: **MDI+: A flexible feature importance framework for random forests**
Presenter: **Tiffany Tang**, University of California, Berkeley, United States
Co-authors: Abhineet Agarwal, Ana Kenney, Yan Shuo Tan, Bin Yu

The mean decrease in impurity (MDI) is commonly used to evaluate feature importances in random forests (RF). It is shown that the MDI for a feature in each fitted tree in an RF is the unnormalized r-squared value in a linear regression of the response on the collection of local decision stumps corresponding to nodes that split on this feature. Building upon this r-squared interpretation of MDI, MDI+ is developed, which generalizes MDI and provides a flexible framework for computing feature importances using RFs. This MDI+ framework is based on a new predictive model, RF+, that allows the analyst to (1) replace the linear regression model and/or r-squared metric with regularized generalized linear models (GLMs) and metrics better suited for the given data structure and (2) incorporate additional features or knowledge to mitigate known biases of decision trees such as their inefficiency in fitting additive or smooth models. Extensive data-inspired simulations show that MDI+ significantly outperforms popular feature importance measures in ranking and identifying relevant features across various settings. Then, in a real-world case study on drug response prediction, MDI+ extracts well-established predictive genes with greater stability and robustness compared to existing feature importance measures. Finally, possible extensions are discussed, and cases of MDI+ are used for extracting interpretable insights from causal forests and heterogeneous treatment effect estimation.

E0514: **Prediction intervals with high dimensional models: With applications in LASSO and deep neural networks**
Presenter: **Zhe Fei**, UC Riverside, United States

A new approach is presented for making inferences about the prediction of continuous outcomes in high-dimensional settings where the number of features greatly exceeds the number of observations. Our method involves repeated applications of the Lasso procedure on the resampled data, treating the resulting smooth learner as a U-statistic. The theoretical properties of the smooth learner are established, and a consistent variance estimator is developed to quantify prediction uncertainty. Our approach is extended to deep neural networks, and the prediction intervals are derived. To demonstrate the effectiveness of our method, simulations have been conducted and applied them to a real-world dataset to predict the DNA methylation age of patients with different tissue samples, which may adequately characterize the ageing process.

EO227 Room 503 ADVANCES IN BAYESIAN THEORY AND COMPUTING Chair: Antonio Lijoi

E0414: **Conditional partial exchangeability: A probabilistic framework for multi-view clustering**
Presenter: **Beatrice Franzolini**, Bocconi University, Italy
Co-authors: Maria De Iorio

Standard clustering techniques assume a common configuration for all features in a dataset. However, when dealing with multi-view or longitudinal data, clusters' shapes and definitions may need to vary across features to accurately describe structure and heterogeneity. This framework requires accounting for within-subject dependence across multiple features, even with different support spaces. Unfortunately, classical model-based clustering techniques fail to account for this dependence. A wide class of Bayesian clustering models designed is introduced to tackle this challenge. The core feature is conditional partial exchangeability, a novel probabilistic paradigm that handles the problem, induces dependence among partitions, ensures tractability, and could ultimately provide a probabilistic framework for the development and study of dependent random partitions of the same subjects.

E0486: **Dependent nonparametric priors based on finite point processes**
Presenter: **Federico Camerlenghi**, University of Milano-Bicocca, Italy
Co-authors: Alessandro Colombi, Raffaele Argiento, Lucia Paci

During the last decade, the Bayesian nonparametric community has focused on defining and investigating prior distributions in the presence of multiple-sample information. A large variety of available models are typically defined by relying on suitable transformations of infinite point processes. A vector of dependent random probability measures is defined for data organized in groups by normalizing a class of dependent finite point processes. It is assumed that the random probability measures to share the same atoms but with different weights to allow the borrowing of information across diverse groups. The model's theoretical properties are studied, i.e., the predictive, posterior and marginal distributions. The random vector of probability measures we propose is then used as a latent structure to define a level-dependent mixture model for clustering with a prior on the number of components. The usefulness of our proposal is also showcased to address extrapolation problems in the presence of multiple populations of species with unknown proportions. In such a setting, closed-form expressions are derived for many statistics of interest, which are still missing in the species literature.

E0797: **Adaptive variational Bayes**
Presenter: **Ilsang Ohn**, Inha University, Korea, South

The adaptive inference is considered based on variational Bayes. Although several studies have been conducted to analyze the contraction properties of variational posteriors, there is still a lack of a general and computationally tractable variational Bayes method that performs adaptive inference. To fill this gap, a novel adaptive variational Bayes framework, which can operate on a collection of models is proposed. The proposed framework first computes a variational posterior over each individual model separately and then combines them with certain weights to produce a variational posterior over the entire model. It turns out that this combined variational posterior is the closest member to the posterior over the entire model in a predefined family of approximating distributions. It is shown that the adaptive variational Bayes attains optimal contraction rates adaptively under very general conditions. In addition, a methodology is provided to maintain the tractability and adaptive optimality of the adaptive variational Bayes even in the presence of an enormous number of individual models, such as sparse models. The general results with several examples are illustrated and new and adaptive inference results are derived.

E0865: **Modeling exchangeable sequences by mixture of mixture and its application**
Presenter: **Shuhei Mano**, The Institute of Statistical Mathematics, Japan

Exchangeable sequences modeled by a mixture of iid random variables following a finite mixture of random probability measures were studied. The results were applied to the problem of simulating outputs of random quantum circuits, which is the central issue in quantum supremacy demonstration. It was demonstrated that the Chinese restaurant metaphor of the proposed modeling simulates the outputs on a classic computer in a reasonable time.

E0870: **Bayesian spatio-temporal clustering of functional data**
Presenter: **Shonosuke Sugasawa**, Keio University, Japan
Co-authors: Tomoya Wakayama, Genya Kobayashi

The use of Bayesian nonparametric modelling and clustering for spatio-temporal functional data is explored. The approach extends a random partition distribution to include spatial similarity, designed explicitly for spatio-temporal scenarios. Efficient algorithms have also been developed to generate posterior samples for this model, making it scalable for large datasets. Simulation studies were conducted and applied to real population data from Tokyo to demonstrate the effectiveness of the method.

EO193 Room 506 RECENT ADVANCES AT THE INTERSECTION OF STATISTICS AND MACHINE LEARNING Chair: Yichen Cheng

E0310: **Analytic natural gradient updates for Cholesky factor in Gaussian variational approximation**
Presenter: **Siew Li Linda Tan**, National University of Singapore, Singapore

Stochastic gradient methods have enabled variational inference for high-dimensional models. However, the steepest ascent direction in the parameter space of a statistical model is actually given by the natural gradient, which premultiplies the widely used Euclidean gradient by the inverse Fisher information. The use of natural gradients can improve convergence, but inverting the Fisher information matrix is daunting in high dimensions. In Gaussian variational approximation, natural gradient updates of the mean and precision of the normal distribution can be derived analytically but do not ensure that the precision matrix remains positively definite. To tackle this issue, the Cholesky decomposition of the covariance or precision matrix is considered, and analytic natural gradient updates of the Cholesky factor are derived, which depend on either the first or second derivative of the log posterior density. Efficient natural gradient updates of the Cholesky factor are also derived under sparsity constraints representing different posterior correlation structures. As Adam's adaptive learning rate does not work well with natural gradients, stochastic normalized natural gradient ascent is proposed with momentum. The efficiency of the proposed methods is demonstrated using logistic regression and generalized linear mixed models.

E0407: **Predictive modeling of transcription-wide association studies via statistical learning methods**
Presenter: **Min Chen**, University of Texas at Dallas, United States

Traditional genomic and transcriptomic predictive models usually require strong assumptions on model structures and data distributions. In contrast, statistical learning models, developed mainly for the purpose of prediction, are less restrictive because they are able to learn with little model assumptions. This is very powerful because it allows for detection without specifying explicitly, e.g., whether the phenotype has additive/multiplicative, dominant/recessive, or epistatic effects. In addition, they can capture nonlinear structures defined by complex genomic and epigenomic regulatory networks. A statistical learning model is proposed to incorporate known regulatory relationships, pathways and epigenomic networks. The model is applied to GTEx and Geuvadis data to improve the prediction of risk and mRNA expression associated with SNPs.

E0354: **Wasserstein Gaussianization and efficient variational Bayes for robust Bayesian synthetic likelihood**
Presenter: **Nhat Minh Nguyen**, The University of Sydney, Australia
Co-authors: Minh-Ngoc Tran, David Nott, Christopher Drovandi

The efficiency of the Bayesian Synthetic Likelihood (BSL) method relies on the normality assumption of the summary statistics. A Wasserstein gradient flow is proposed to be used to approximately transform the distribution of the summary statistics into a Gaussian distribution. BSL also implicitly requires compatibility between simulated summary statistics under the working model and the observed summary statistics. This requirement has been facilitated by the robust BSL method developed recently in the literature. The Wasserstein Gaussianing transformation is combined with robust BSL, together with an efficient Variational Bayes procedure for posterior approximation, to develop a highly efficient and reliable approximate Bayesian inference method for likelihood-free problems.

E0678: **A Bayesian semi-supervised approach to keyphrase extraction**
Presenter: **Yichen Cheng**, Georgia State University, United States

In the era of big data, people are benefited from the existence of tremendous amounts of information. However, the availability of said information may pose great challenges. For instance, one big challenge is how to extract useful yet succinct information in an automated fashion. As one of the first few efforts, keyword extraction methods summarize an article by identifying a list of keywords. Many existing keyword extraction methods focus on the unsupervised setting, with all keywords assumed unknown. In reality, a (small) subset of the keywords may be available for a particular article. A rigorous probabilistic model based on a semi-supervised setup is proposed to utilize such information. The method incorporates the graph-based information of an article into a Bayesian framework via an informative prior so that our model facilitates formal statistical inference, which is often absent from existing methods. Both Markov-chain Monte Carlo algorithms based on Gibbs samplers and Variational Bayesian methods are developed to overcome the difficulty arising from high-dimensional posterior sampling. A false discovery rate (FDR) based approach is employed for selecting the number of keywords, while the existing methods use ad-hoc threshold values. The numerical results show that the proposed method compared favourably with state-of-the-art methods for keyword extraction.

E1048: **Convergence results of numerically estimated JKO scheme**
Presenter: **Paco Tseng**, University of Sydney, Australia
Co-authors: Minh-Ngoc Tran

Applications of Wasserstein gradient flow in the field of Bayesian Computation have been a trending topic. The iterative scheme developed under this framework is termed JKOscheme, and the theory states that the solution sequence of the JKO scheme converges to the gradient flow of the KL divergence with respect to the posterior distribution. At each iteration, the solution is determined by a vector field. In particular, the convergence results rely on the tractable vector field. However, in practice, the vector field is often estimated by some numerical methods for various reasons, for instance, large data sets and analytically intractable vector fields. The convergence results of the JKO scheme, when the estimated vector field is in place, are provided.

EO128 Room 603 APPLIED MATHEMATICS, STATISTICS, AND AI IN PORTFOLIO OPTIMIZATION Chair: Chi Seng Pun

E0508: **The constrained Dantzig-type estimator: An application to selection of high-dimensional portfolios**
Presenter: **Dechuan Zhu**, Nanyang Technology University, Singapore
Co-authors: Chi Seng Pun

Asset selection problems, especially sparse portfolio construction, have been getting uprising attention in recent years due to the vast increase in available assets, paired with comparably limited yet noisy information in the market. Moreover, investment activities are often restricted by various constraints, e.g. budget constraints. A constrained Dantzig-type estimator (CDE) is developed for sparse learning problems with equality constraints, such as sparse portfolio construction. It is shown that CDE is able to produce an estimate of the oracle solution and counter the curse of dimensionality. At the same time, its non-asymptotic statistical error bounds under l1 and l2 norms are derived. Compared to the constrained Lasso, CDE has a significant computational advantage as CDE can be obtained via the solution to a linear problem. Moreover, CDE is extremely versatile and widely applicable. Extensive simulations and empirical studies show that sparse portfolios constructed using CDE have superior out-of-sample performance compared to various benchmark portfolios, including the equally weighted portfolio.

E0529: **Equilibrium control under cumulative prospect theory via reinforcement learning**
Presenter: **Nixie Sapphira Lesmana**, Nanyang Technological University, Singapore

Cumulative prospect theory (CPT) optimization extends the well-known expected utility maximization by distorting probabilities and incorporating S-shaped utility functions to capture human decisions or preferences better. CPT optimization is investigated from the lens of reinforcement learning (RL). Firstly, noting that CPT can be rewritten as a function of quantiles, selected distributional RL techniques are reviewed that deal with the prediction of quantile functions and range from tabular to deep function approximation and highlight both their applicability and limitations for CPT Q-function prediction. Secondly, noting the time inconsistency introduced by probability distortions, it is shown that these RL algorithms aim to learn the subgame-perfect equilibrium (SPE) policy class. Drawing on these two perspectives, a novel RL algorithm is proposed that can learn a (deep, approximate) SPE policy under CPT. The performance of our algorithm is empirically tested on several variant environments of casino gambling.

E0624: **Sensitivity of robust optimization problems with ambiguity on semimartingale differential characteristics**
Presenter: **Kyunghyun Park**, Nanyang Technological University, Singapore
Co-authors: Daniel Bartl, Ariel Neufeld

A sensitivity analysis is provided for robust optimization problems, where model ambiguity is captured by a closed ball (with respect to some suitable norms) around each semimartingale differential characteristic of a postulated reference Ito-semimartingale. Assuming a decision maker seeks to derive a robust control such that its stochastic integral optimizes her minimax value function, the first-order correction, which is defined as the first-order derivative of the minimax value function with respect to the radius of the ball at 0, is obtained. In particular, the correction is characterized in terms of the optimizer of the value function under the postulated reference semimartingale without model ambiguity and the gradient of the function at the optimum. The approach relies on dual norm representations and the tools of backward stochastic differential equations. In the context of finance and economics, some possible applications and extensions are discussed within the proposed results.

E0917: **Portfolio selection based on anomaly detection using GANs**
Presenter: **Yongjae Lee**, Ulsan National Institute of Science and Technology, Korea, South

The application of Generative Adversarial Networks (GANs) for anomaly detection in stock time series data is explored, and subsequently, employ this approach for portfolio selection across industry sectors. Recently, generative models have garnered substantial interest in the realm of artificial intelligence, with GANs standing out due to their distinctive dual architecture, comprising a generator that creates synthetic data and a discriminator that differentiates between genuine and fake data. Capitalizing on this feature, various GAN-based anomaly detection models have emerged. Our research introduces a novel method for devising an industry sector portfolio grounded on the anomaly detection signals derived from GANs applied to stock time series data. The GANs are specifically trained for each industry sector, generating anomaly signals that are consolidated to assemble the final portfolio. This approach enables the intuitive identification of sectors facing challenges and facilitates the adjustment of the portfolio in response, ultimately offering enhanced explainability to investors.

E1270: **Understanding the difficulty of achieving dynamic optimality in time inconsistent problems**
Presenter: **Jingxiang Tang**, Nanyang Technological University, Singapore

The variance of cumulative rewards arises naturally as part of the decision-making criterion in much important reinforcement learning (RL) applications, such as portfolio and resource allocations. The time inconsistency induced by such a criterion makes the search for globally-optimal policy difficult. Many proposals have been made to resolve this, with episodic policy gradient (EPG) as one popular method. This paper highlights the difficulties of actually attaining global optimality with EPG and introduces alternative optimality: subgame perfect equilibrium (SPE) that is achievable in RL. Both optimality types on portfolio optimization and optimal execution problems in finance are empirically evaluated. Our results suggest that there are some instances where EPG does not learn the desired globally optimal policy while SPE provides a better solution.

EO030 Room 604 RECENT ADVANCES IN TIME SERIES ECONOMETRICS Chair: Kaiji Motegi

E0490: **Time-varying ambiguity shocks and business cycles**
Presenter: **Xiaojing Cai**, Okayama University, Japan

Relationships between ambiguity and macroeconomic variables are investigated. Following previous research, ambiguity is measured as a weighted average of the variances of probabilities. A longer dataset is employed, and whether ambiguity impacts key macroeconomic variables such as outputs and inflation in the U.S. over the post-World War II period is explored. In addition to ambiguity obtained from market risk premiums, it is estimated from size, value, and momentum risk premiums. To assess the effects of ambiguity over 70 years, a time-varying parameter vector autoregressive model with stochastic volatility (TVP-VAR-SV) that allows us to reflect structural changes in the economy is employed. The results provide evidence that the relationships between ambiguity and macroeconomic variables vary over time. Specifically, it is found that an increase in ambiguity led to an increase in output during the high inflation periods in the 1970s and the 1980s, which is consistent with the ambiguity lover behaviour previously. A negative relationship between ambiguity and inflation in the 1950s and a positive relationship in the 2000s are also observed, which indicates that unfavourable outcomes differ between the former and the latter period. Moreover, it is uncovered that ambiguity obtained from other risk premiums is weakly associated with macroeconomic variables, suggesting that size and value risk factors do not capture information about the entire market.

E0587: **Can NFTs risk hedge other traditional assets after the COVID19 pandemic?**
Presenter: **Wenting Zhang**, Kobe University, Japan

The aim is to analyze the dynamic spillover effects between the NFT market, Bitcoin market, oil market, gold futures market, S&P 500 stock index, bond market, and US dollar index over three time periods: the full sample period before the COVID-19 outbreak, and after the COVID-19 outbreak by employing the DCC-GARCH-based connectedness model. Furthermore, the DCC-GARCH-t-Copula model, Risk Parity Portfolio (RPP) model, and Minimum Connectedness Portfolio (MCoP) model are applied to evaluate the bilateral dynamic hedge ratios, portfolio weights, and the multivariate portfolio performance of these financial assets, respectively. The findings suggest that NFTs are volatility spillover transmitters during either period and that the COVID-19 outbreak accelerates the speed and intensity of volatility spillovers from NFTs to traditional financial markets. Moreover, most of the volatility spillovers from NFTs are caused by endogenous shocks, which implies that NFTs can prevent financial risk contagion. The results of risk hedging analysis and multivariate portfolio results show that NFTs can effectively hedge other traditional financial assets as long positions. NFTs can also reduce investment risk despite their smaller weighting in a multivariate portfolio. The explosion of COVID-19 makes MCoP slightly outperform RPP, although RPP outperforms MCoP in general.

E0547: **Comparing factor models with conditioning information**
Presenter: **Seok Young Hong**, Lancaster University Management School, United Kingdom

A novel framework is developed to conduct asymptotically valid tests for comparing factor models with conditioning information. The tests are based on a metric analogous to the squared Sharpe ratio improvement measure that is used to gauge the extent of model mispricing in an unconditional setting. An estimator for the metric is proposed, and its limiting properties, establishing the asymptotic normality, are studied. An advantage of our framework is that it can be applied without an a priori knowledge of the persistent nature of the conditioning variables. A range of dependence classes is accommodated, including stationary, nearly stationary, integrated, and local-to-unity.

E0341: **Estimating high-dimensional Markov-switching VARs**
Presenter: **Kenwin Maung**, University of Rochester, United States

Maximum likelihood estimation of large Markov-switching vector autoregressions (MS-VARs) can be challenging or infeasible due to parameter proliferation. A sparse framework is adopted to accommodate situations where dimensionality may be of comparable order to or exceeds the sample size. Two penalized maximum likelihood estimators are proposed with the Lasso or the smoothly clipped absolute deviation (SCAD) penalty. It is shown that both estimators are estimation consistent, while the SCAD estimator also selects relevant parameters with probability approaching one. A modified EM algorithm is developed for the case of Gaussian errors, and simulations show that the algorithm exhibits desirable finite sample performance. In applying short-horizon return predictability in the US, a 15-variable 2-state MS-VAR(1) and obtaining the often reported counter-cyclicality in predictability are estimated. The estimators' variable selection property helps identify predictors that contribute strongly to predictability during economic contractions but are otherwise irrelevant in expansions. Furthermore, out-of-sample analyses indicate that large MS-VARs can significantly outperform "hard-to-beat" predictors like the historical average.

E0169: **Midastar: Threshold autoregression with data sampled at mixed frequencies**
Presenter: **Kaiji Motegi**, Kobe University, Japan
Co-authors: John Dennis

The aim is to propose Midastar models by combining the mixed data sampling (MIDAS) approach and the threshold autoregressive (TAR) models. The Midastar model of the first kind is designed for a low-frequency target variable and a high-frequency threshold variable, while the second kind is designed for the reverse case. The Midastar models can detect threshold effects accurately, while the aggregated TAR model has a risk of finding spurious non-threshold effects. The Midastar models have desired asymptotic and finite sample properties. As an empirical application, the Midastar model of the first kind is fitted to Japan's COVID-19 data. The target variable is the growth of weekly hospitalization, and the threshold variable is the growth of daily new confirmed cases in Japan. Statistically significant threshold effects, revealing heterogeneity between the contraction and expansion regimes of the pandemic, are detected. The threshold effects vanish once the daily new confirmed cases are aggregated into the weekly level, a spurious non-threshold effect.

EO216 Room 605 STOCHASTIC FRONTIER AND PRODUCTIVITY ANALYSIS WITH PANEL DATA APPLICATIONS Chair: Kai Sun

E0268: **A single-index smooth-coefficient stochastic frontier model**
Presenter: **Kai Sun**, Shanghai University, China
Co-authors: Subal Kumbhakar

A single-index semiparametric smooth-coefficient stochastic production frontier model where technology parameters and technical inefficiency are unknown smooth functions of production environmental variables is considered. The variables affecting the technology parameters and technical inefficiency are allowed to be either different or the same. Marginal effect formulae of these variables on output and inefficiency are provided. Output growth is decomposed into technical change (TC), input-driven component, and efficiency change (EC), while TFP growth is decomposed into TC, scale component, and EC. Simulations show that our estimation procedures work quite well in finite samples. Finally, using firm-level datasets, the methodology is applied to study the productivity and efficiency, along with their temporal behaviours, for the two Norwegian industry groups: Knowledge Intensive Business Services and high-technology manufacturing industries, and find significant impacts of geographical industrial concentration, export intensity, and location on firm's technology frontier and output.

E0544: **A semiparametric stochastic frontier model with two-way fixed effects and nonparametric inefficiency function**
Presenter: **Taining Wang**, Capital University of Economics and Business, China
Co-authors: Kai Sun

A semiparametric stochastic frontier panel model is proposed, relaxing conventional parametric assumptions on both inefficiency and frontier. First, distributional assumptions are not imposed on the inefficiency for identification, but only the existence of its mean conditioning on observables is assumed. Unlike existing models, the level of inefficiency mean function is identified by estimating it with the frontier in a sequential step, achieved by applying conventional two-way within the transformation. Second, to combat the curse of dimensionality, a single-index structure in the inefficiency mean and the inputs elasticity is introduced, which can vary with contextual environmental variables in a nonlinear fashion. Third, the inefficiency is disentangled from latent heterogeneities in firm and time dimensions. A three-step estimation procedure that combines the use of series and kernel estimator is employed, and their appealing finite-sample performance through simulation studies is demonstrated. Our model's applicability is showcased by performing an efficiency analysis in the banking industry.

E0926: **Bank efficiency and credit risk: Evidence from the commercial banks in China**
Presenter: **Kai Du**, University Of Wollongong, Australia
Co-authors: Kai Sun

Using commercial banks in China, the effect of credit risk is investigated on the production frontier and technical efficiency using a semi-parametric model of stochastic frontier analysis. The empirical results reveal that an increase in loan loss provision, as a proportion of the total assets, increases the labour productivity of a commercial bank but the increase in equity as a proportion of the total assets may not have a similar effect. In other words, an increase in the loan loss provision suggests that the quality of loans has worsened, but an increase in the equity ratio simply represents a decrease in financial leverage.

E1116: **Leveraging innovation for improved service productivity: Insights from endogenous stochastic frontier analysis**
Presenter: **Fikru Kefyalew Alemayehu**, Inland Norway University of Applied Sciences, Norway
Co-authors: Subal Kumbhakar, Kai Sun

In recent years, there has been a surge of interest in the relationship between innovation, service productivity, and efficiency. Despite increased research attention, this relationship is still poorly understood. To address this issue, a recent instrumental variable approach of Endogenous Stochastic Frontier Analysis is utilized within a semi-parametric framework based on data from the Community Innovation Survey of Norwegian firms from 2002 to 2016. This method is essential for dealing with the issue using variables such as R&D spending and collaboration as instruments for innovation. The findings show a significant positive relationship between service productivity and innovation. Overall, the aim is to bridge the gap in this research area by providing useful insights into the role of innovation in service productivity.

E1188: **Revisiting the public capital productivity puzzle**
Presenter: **Zhezhi Hou**, Southwestern University of Finance and Economics, China

Although public capital is crucial in production and economic growth, most empirical studies that assume Cobb-Douglas production technology find that the estimated returns of the public capital are either negative or statistically insignificant when fixed effects are controlled. It is hypothesized that these counterintuitive estimates may be due to restrictive functional forms and/or ignoring cross-sectional dependence in the estimation of the production technology. To investigate this hypothesis, several semi/nonparametric models are deployed with fixed effects and/or multi-factor error structures to reexamine the impact of public capital on state GDP in the U.S. from 1970 to 1986. After going through a battery of models, positive, neutral effects, negative non-neutral effects, and heterogeneous overall effects of public capital on output are found. The heterogeneity helps explain the negative or insignificant effects estimated from a constant elasticity parametric model, which captures only the mean/median effect. It is also found that controlling cross-sectional dependence tends to increase the above estimates. In addition, when public capital is disaggregated into its components, positive effects for the water and sewer systems, mixed effects for the highways, and negative effects for other buildings are found.

E1332: **Oil shock, policy uncertainty and stock return: An analysis based on the Bart method**
Presenter: **Jianhua Zhou**, Sun Yat-Sen university, China

An extension of the time-varying network dependence panel (TNDP) mode that incorporates the identification of dominant units is utilized. The objective is to re-examine the impacts of the oil shock and policy uncertainty on stock returns in a network framework with heterogeneity both over time and across sections. Bayesian econometric methods using additive regression trees are employed and are used to address extreme observations. We argue that regression tree models are well-suited for macroeconomic nowcasting due to their flexibility and ability to model outliers.

EO149 Room 606 RECENT ADVANCES IN TIME SERIES AND CHANGE-POINT ANALYSIS Chair: Chun Yip Yau

E0426: **GARCH-type factor model**
Presenter: **Yuanbo Li**, University of International Business and Economics, China
Co-authors: Chi Tim Ng, Chun Yip Yau

A new model is proposed for factor analysis of multivariate time series. The latent factors in the model are linked to observed time series through a deterministic relationship in a manner that is similar to the volatility process of the GARCH model. Mathematically-tractable quasi-likelihood is constructed for the proposed GARCH-type factor model, allowing efficient statistical inference even for high-dimensional time series incorporating non-Gaussian idiosyncratic components. Asymptotic theory for statistical inference of the proposed model is also developed. The applicability of the proposed model to real data is demonstrated through macroeconomic data and forward rate data of bonds. The factors extracted are then utilized to elucidate the risk premium of U.S. government bonds.

E0448: **Scalable semiparametric spatio-temporal regression for large data analysis**
Presenter: **Ting Fung Ma**, University of South Carolina, United States
Co-authors: Fangfang Wang, Jun Zhu, Anthony Ives, Katarzyna Lewinska

With the rapid advances in data acquisition techniques, spatio-temporal data are becoming increasingly abundant in a diverse array of disciplines. Spatio-temporal regression methodology is developed for analyzing large amounts of spatially referenced data collected over time, motivated by environmental studies utilizing remotely sensed satellite data. In particular, a semiparametric autoregressive model without the usual Gaussian assumption and devise a computationally scalable procedure that enables the regression analysis of large datasets is specified. The model parameters by maximum pseudolikelihood are estimated, and it is shown that the computational complexity can be reduced from cubic to linear of the sample size. Asymptotic properties under suitable regularity conditions are further established that inform the computational procedure to be efficient and scalable. A simulation study is conducted to evaluate the finite-sample properties of the parameter estimation and statistical inference. The methodology is illustrated by a dataset with over 2.96 million observations of annual land surface temperature, and a comparison with an existing state-of-the-art approach to spatiotemporal regression highlights the advantages of this method.

E0513: **Inference for multiple change-points in piecewise locally stationary time series**
Presenter: **Wai Leong Ng**, The Hang Seng University of Hong Kong, Hong Kong

In a wide range of applications, such as econometrics, finance, and seismology, the stochastic properties of the observed time series change over time, and this phenomenon can be modelled by locally stationary time series models with time-varying parameter curves. However, the assumption of local stationarity may be violated. For example, abrupt changes in parameter values and the slope of the parameter curves may occur at some time points, referred to as jump points and kink points, respectively. In this case, piecewise, locally stationary time series models are more appropriate for modelling the stochastic properties of the time series. In contrast to the classical change-point setting in time series, methods for detecting both jumps and kinks in a piecewise locally stationary time series model are less developed. A three-step criterion-based procedure is presented for multiple change-point inferences in a piecewise locally stationary time series with possible jumps and kinks in its parameter curve. Theoretical properties are established, including consistency of the number and locations of the change-point estimation and the asymptotic exactness of the confidence intervals. Simulation studies and real data applications are provided to illustrate the performance of the proposed method.

E0586: **Weighted kernel estimators for forecasting under breaks**
Presenter: **Sze Him Isaac Leung**, The Chinese University of Hong Kong, Hong Kong

Predicting future observations in non-parametric regression models that are subject to a structural break at an unknown time point is studied. A weighted kernel estimator is proposed to estimate the post-break regression function and forecast future observations, where the weights are location- and time-dependent. It is shown that putting some weights to pre-break observations can improve the estimate of the post-break regression function in terms of the mean squared forecast error (MSFE). This is related to the bias-variance trade-off induced by including pre-break observations. The MSFE optimal weights and bandwidth are estimated in a two-step approach, and the properties are examined. Simulation studies demonstrate that the proposed weighted kernel estimator has a smaller MSFE than post-break methods under various non-parametric regression functions.

E0699: **Generalized multivariate threshold autoregressive models with linearly partitioned threshold space**
Presenter: **Gan Yuan**, Columbia University, United States

A k-dimensional multiple-regime vector threshold autoregressive model is considered, in which the regime-switching mechanism is governed by another bivariate observable time series, known as threshold variables. Specifically, the regimes are induced by a partition of the threshold space by an unknown number of threshold lines. The process is governed by a specific vector autoregressive (VAR) model within each regime. The model selection and parameter estimation are formulated into a minimization problem based on the Minimum Description Length (MDL) principle, and the number of threshold lines, parametric forms of threshold lines and VAR model parameters in each regime simultaneously are estimated. Theoretically, it is shown that the MDL estimators of threshold lines are n-consistent, and their weak convergence is characterized. The main novelty in the proof is introducing a new functional space $\mathbb{G}$ for the local MDL difference functions, as opposed to earlier works, in which the weak convergence was established in the classical $\mathbb{D}$ space. Finally, some empirical studies are conducted with simulated datasets and real data analysis on US interest rates is performed.

EO177 Room 702 RECENT DEVELOPMENT IN NETWORK DATA ANALYSIS Chair: Yichuan Zhao

E0388: **Supervised centrality via sparse spatial autoregression**
Presenter: **Yingying Ma**, Beihang University, China
Co-authors: Wei Lan, Chenlei Leng, Ting Li, Hansheng Wang

The social characteristics of the players in a social network are closely associated with their network positions. Identifying the influential players in a network is of great importance as it helps to understand how ties are formed and how information is propagated, and in turn, can guide the dissemination of new information. A new notion of supervised centrality is proposed, emphasizing that the centrality of a player is task-specific. A novel sparse spatial autoregression is developed by introducing individual heterogeneity to each user to estimate the supervised centrality and identify important players. To overcome the computational difficulties in fitting the model for large social networks, a forward-addition algorithm is further developed, and it is shown that it can consistently identify a superset of the influential nodes. The method is applied to analyze three responses in Henan Floods data: the number of comments, reposts and likes, and obtain meaningful results. A simulation study further corroborates the developed theory.

E0438: **Spectral analysis on networks with covariates**
Presenter: **Wanjie Wang**, National University of Singapore, Singapore

Social network data record the connections between objects. In past decades, the study of social network data has been an important topic. Together with the observed connections, the profile of the object itself is usually also recorded. The data on the node-level profile covariates is called. Both the covariates and the network reflect the underlying data structure. The results on how the covariates will help to solve problems on the networks, such as community detection, and how the network helps with problems about the covariates, such as feature selection, are introduced. Spectral methods, which are computationally efficient in such problems, are mainly considered. Hence, the theoretical results on the control of the eigenvectors/eigenvalues are introduced when both the networks and covariates are combined. Such results are useful for further studies.

E0451: **Higher-order accurate two-sample network inference and network hashing**
Presenter: **Dong Xia**, Hong Kong University of Science and Technology, Hong Kong

Two-sample hypothesis testing for comparing two networks is an important yet difficult problem. Major challenges include: potentially different sizes and sparsity levels; non-repeated observations of adjacency matrices; computational scalability; and theoretical investigations, especially on finite-sample accuracy and minimax optimality. The first provably higher-order accurate two-sample inference method is proposed by comparing network moments. The method extends the classical two-sample t-test to the network setting. We make weak modelling assumptions and can effectively handle networks of different sizes and sparsity levels. We establish strong finite-sample theoretical guarantees, including rate-optimality properties. Our method is easy to implement and computes fast. We also devise a novel nonparametric framework of offline hashing and fast querying, particularly effective for maintaining and querying very large network databases. The effectiveness of the method by comprehensive simulations. The method is applied to two real-world data sets and discovers interesting novel structures.

E0558: **Testing stochastic block models via the maximum sampling entry-wise deviation**
Presenter: **Wei Lan**, Southwestern University of Finance and Economics, China

The stochastic block model (SBM) has been widely used to analyze network data. Various goodness-of-fit tests have been proposed to assess the adequacy of model structures. To the best of our knowledge, however, none of the existing approaches is applicable for sparse networks in which the connection probability of any two communities is of order $O(n^{-1}\log n)$, and the number of communities is divergent. A novel goodness-of-fit test is proposed for the stochastic block model to fill this gap. The key idea is combining a previous test concept with a sampling process that alleviates the negative impacts of network sparsity. Theoretically, the proposed test statistic converges is demonstrated to the Type-I extreme value distribution under the null hypothesis, regardless of the network structure. Accordingly, it can be applied to both dense and sparse networks. In addition, the asymptotic power against alternatives is obtained. Moreover, a bootstrap corrected test statistic is introduced to improve the finite sample performance, recommend an augmented test statistic to increase the power and extend the proposed test to the degree-corrected SBM. Simulation studies and two empirical examples with dense and sparse networks indicate that the proposed method performs well.

E0885: **Variational inference: Posterior threshold improves network clustering accuracy in sparse regimes**
Presenter: **Can Minh Le**, University of California, Davis, United States

The variational inference has been widely used in machine learning literature to fit various Bayesian models. This method has been successfully applied in network analysis to solve community detection problems. Although these results are promising, their theoretical support is only for relatively dense networks, an assumption that may not hold for real networks. In addition, it has been shown recently that the variational loss surface has many saddle points, which may severely affect its performance, especially when applied to sparse networks. The aim is to propose a simple way to improve the variational inference method by hard thresholding the posterior of the community assignment after each iteration. Using a random initialization that correlates with the true community assignment, it is shown that the proposed method converges and can accurately recover the true community labels, even when the average node degree of the network is bounded. The extensive numerical study further confirms the advantage of the proposed method over classical variational inference and another state-of-the-art algorithm.

EO107 Room 703 TRUSTWORTHY AND EFFICIENT MACHINE LEARNING Chair: Yao Li

E0332: **Trusted aggregation (TAG): Model filtering backdoor defense in federated learning**
Presenter: **Yao Li**, University of North Carolina at Chapel Hill, United States
Co-authors: Joseph Lavond, Minhao Cheng

Federated learning is a framework for training machine learning models from multiple local data sets without access to the data in the aggregate. A shared model is jointly learned through an interactive process between the server and clients that combines locally known model gradients or weights. However, the lack of data transparency naturally raises concerns about model security. Recently, several state-of-the-art backdoor attacks have been proposed, which achieve high attack success rates while simultaneously being difficult to detect, leading to compromised federated learning models. Motivated by differences in the outputs of models trained with and without the presence of backdoor attacks, a defense method that can prevent backdoor attacks from influencing the model while maintaining the accuracy of the original classification task is proposed. TAG leverages a small validation data set to estimate the largest change a benign user's local training can make to the shared model, which can be used as a cutoff for returning user models. Experimental results on multiple data sets show that TAG defends against backdoor attacks even when 40% of the user submissions to update the shared model are malicious.

E0333: **Learning manifold-structured data using deep neural networks: Theory and applications**
Presenter: **Rongjie Lai**, Rensselaer Polytechnic Institute, United States

Deep artificial neural networks have succeeded greatly in many problems in science and engineering. Our recent efforts are discussed to develop DNNs capable of learning non-trivial geometry information hidden in data. The first part discusses work on advocating using a multi-chart latent space for better data representation. Inspired by differential geometry, a Chart Auto-Encoder (CAE) is proposed, and a universal approximation theorem on its representation capability is proved. CAE admits desirable manifold properties that conventional auto-encoders with a flat latent space fail to obey. Statistical guarantees on the generalization error are further established for trained CAE models, and their robustness is shown to be noise. The numerical experiments also demonstrate satisfactory performance on synthetical and real-world data.

E0343: **Barycenter estimation of positive semi-definite matrices with Bures-Wasserstein distance**
Presenter: **Jingyi Zheng**, Auburn University, United States
Co-authors: Huajun Huang, Yuyan Yi, Yuexin Li, Shu-Chin Lin

Brain-computer interface (BCI) builds a bridge between the human brain and external devices by recording brain signals and translating them into commands for devices to perform the user's imagined action. The core of the BCI system is the classifier that labels the input signals as the user's imagined action. The classifiers that directly classify covariance matrices using Riemannian geometry are widely used not only in the BCI domain but also in a variety of fields. However, the existing Affine-Invariant Riemannian-based methods suffer from issues such as being time-consuming, not robust, and having convergence issues when the dimension and number of covariance matrices become large. To address these challenges, the mathematical foundation is established for the Bures-Wasserstein distance and new algorithms are proposed to estimate the barycenter of positive semi-definite matrices efficiently and robustly. Both theoretical and computational aspects of Bures-Wasserstein distance and barycenter estimation algorithms are discussed. With extensive simulations, the accuracy, efficiency, and robustness of the barycenter estimation algorithms coupled with the Bures-Wasserstein distance are comprehensively investigated. The results show that Bures-Wasserstein-based barycenter estimation algorithms are more efficient and robust.

E0620: **Improving neural networks interpretability and trustworthiness using polynomials and feature interactions**
Presenter: **Pablo Morala**, Universidad Carlos III de Madrid, Spain
Co-authors: Rosa Lillo, Inaki Ucar, Jenny Alexandra Cifuentes

Assessing the interpretability and explainability of neural networks is a key factor in adopting their use in a wide range of applications, where their black-box nature poses a problem regarding their trustworthiness. The latest advances in NN2Poly will be presented. NN2Poly is a method that obtains a relationship between a given trained neural network and a polynomial using Taylor expansion and combinatorial properties. This method transforms the nonlinearities in each layer and neuron into polynomials, which can be combined into a final polynomial representation of the neural network. This allows using the polynomial coefficients to interpret the importance of features in the global model, including the interactions between variables up to a certain order. This key aspect is not usually present in other explainability tools. Furthermore, this framework can be used to explore the internal learning process of the neural network by analyzing the obtained polynomials at each layer. Simulations will be presented with both synthetic data (where the interactions are known) and the application to real datasets in regression and classification tasks.

E0783: **Learning to make adherence-aware recommendations**
Presenter: **Guanting Chen**, University of North Carolina at Chapel Hill, United States

As AI systems continue to make recommendations for human decision-making, it is frequently observed that human agents sometimes disregard these recommendations. In such cases, it may be beneficial for the AI system to refrain from providing the "optimal" recommendation, which assumes perfect adherence from the agent. A proposed decision-making model considers adherence-aware recommendations, accounting for the varying levels of adherence exhibited by human agents across different states and actions. Aside from the model, accountable and near-optimal reinforcement learning algorithms specifically designed to address adherence-aware recommendations are also introduced.

EO198 Room 704 BAYESIAN STRUCTURE DISCOVERY Chair: Sameer Deshpande

E0368: **The G-Wishart weighted proposal algorithm: Efficient posterior computation for Gaussian graphical models**
Presenter: **Willem van den Boom**, National University of Singapore, Singapore
Co-authors: Alexandros Beskos, Maria De Iorio

Gaussian graphical models can capture complex dependency structures among variables. Bayesian inference is attractive for such models as it provides principled ways to incorporate prior information and quantify uncertainty through the posterior distribution. However, posterior computation under the conjugate G-Wishart prior distribution on the precision matrix is expensive for general non-decomposable graphs. Therefore a new Markov chain Monte Carlo (MCMC) method is proposed named the G-Wishart weighted proposal algorithm (WWA). WWA's distinctive features include delayed acceptance MCMC, Gibbs updates for the precision matrix and an informed proposal distribution on the graph space that enables embarrassingly parallel computations. Compared to existing approaches, WWA reduces the frequency of the relatively expensive sampling from the G-Wishart distribution. This results in faster MCMC convergence, improved MCMC mixing and reduced computing time. Numerical studies on simulated and real data show that WWA provides a more efficient tool for posterior inference than competing state-of-the-art MCMC algorithms.

E0936: **The multivariate spike-and-slab LASSO: Algorithms, asymptotics, and inference**
Presenter: **Sameer Deshpande**, University of Wisconsin–Madison, United States

Multivariate linear regression models are considered to predict q correlated responses (of possibly mixed type) using a common set of p predictors. The interest lies not only in determining whether a particular predictor has a direct or marginal effect on each response but also in understanding the residual dependence between the outcomes. A Bayesian procedure is proposed for such determination using continuous spike-and-slab priors. Rather than relying on a stochastic search through the high-dimensional parameter space, an Expectation Conditional Maximization algorithm targeting modal estimates of the matrix of regression coefficients and the residual precision matrix is developed. A key feature of the method is the model of the uncertainty about which parameters are negligible. Posterior contraction rates are further derived, and several strategies for quantifying posterior uncertainty are discussed.

E0938: **Bayesian modal regression based on mixture distributions**
Presenter: **Ray Bai**, University of South Carolina, United States

Compared to mean regression and quantile regression, the literature on modal regression is very sparse. A unified framework is proposed for Bayesian modal regression based on a family of unimodal distributions indexed by the mode along with other parameters that allow for flexible shapes and tail behaviours. Sufficient conditions for posterior propriety are derived under an improper prior on the mode parameter. Following prior elicitation, regression analysis of simulated data and datasets from several real-life applications are carried out. Besides drawing inferences for covariate effects that are easy to interpret, prediction and model selection under the proposed Bayesian modal regression framework is considered. Evidence from these analyses suggests that the proposed inference procedures are very robust to outliers, enabling one to discover interesting covariate effects missed by mean or median regression and to construct much tighter prediction intervals than those from mean or median regression.

E0941: **Structure learning with global-local prior-penalty dual**
Presenter: **Jyotishka Datta**, Virginia Polytechnic Institute and State University, United States
Co-authors: Anindya Bhadra, Sayantan Banerjee, Ksheera Sagar

High-dimensional data with complex dependence structure is routinely observed in many areas of science and engineering, and the problem of sparse precision matrix estimation is one such methodological problem fundamental for network estimation. Although both Bayesian and frequentist approaches exist, obtaining good Bayesian and frequentist properties under the same prior-penalty dual is difficult, complicating justification. Recent developments in precision matrix estimation will be briefly reviewed using global-local shrinkage priors, the state-of-the-art Bayesian tool for sparse signal recovery. Possible solutions that lead to a prior-penalty dual will be proposed, offering fully Bayesian uncertainty quantification and computationally efficient point estimates. The posterior convergence rate of the precision matrix estimate is established, matching the oracle rate and the frequentist consistency of the posterior mode. Computationally efficient algorithms for both optimization and sampling are developed respectively for obtaining the penalized likelihood and fully Bayesian estimation problems. It is also illustrated with a protein-protein interaction network estimation in B-cell lymphoma, and future directions are pointed out.

E0943: **Identifying interpretable latent structures in factor analysis**
Presenter: **Maoran Xu**, Duke University, United States

Factor models have been widely used to summarize the variability of high-dimensional data through a set of factors with much lower dimensions. Gaussian linear factor models have been particularly popular due to their interpretability and ease of computation. However, many real data violate the multivariate Gaussian assumption. To characterize higher-order dependence and non-linearity, models that include factors as predictors in flexible multivariate regression are popular, with GP-LVMs using Gaussian process (GP) priors for the regression function and VAEs using deep neural networks. Unfortunately, such approaches lack identifiability and interpretability, and tend to produce brittle and non-reproducible results. To address these challenges and simplify nonparametric factor models while maintaining flexibility, we propose the NIFTY framework, which parsimoniously transforms uniform latent variables using 1d non-linear mappings, and then applies a linear generative model. The induced multivariate distribution falls in a flexible class while maintaining simple computation and interpretation. We show identifiability of NIFTY, and empirically study NIFTY with both simulated and real data, observing good performance in dimension reduction in various tasks, including moment estimation and multivariate density estimation.

EO109 Room 705 NOVEL METHODS, ALGORITHMS AND THEORY FOR HIGH-DIMENSIONAL DATA Chair: Qing Mai

E0174: **Robust estimation of central subspace under high-dimensional and elliptical-contoured design**
Presenter: **Jing Zeng**, University of Science and Technology of China, China
Co-authors: Qing Mai

Sufficient dimension reduction (SDR) is a valuable tool to tackle high dimensionality while maintaining the primary information of the prediction problem, and it has demonstrated great promise in many applications. There exists a variety of high-dimensional sufficient dimension reduction methods in the literature. However, they all rely on the sub-Gaussian assumption of the predictors' marginal distribution or conditional distribution. Such a light-tailedness assumption is frequently violated in real life. A new methodology is proposed to estimate the central subspace consistently when the predictor exhibits heavy-tailedness. The novel proposal overcomes the heavy-tailedness issue and the high dimensionality. Under a general regression model assumption and the elliptically-contoured distribution assumption of the predictor, an invariance result between the CS and a surrogate subspace is established. TEstimatingthe surrogate subspace avoids the heavy-tailedness issue and can be implemented using existing high-dimensional SDR methods. Theoretically, the proposal enjoys satisfactory consistency, and the convergence rate is shown to achieve optimality. Empirically, the efficiency and effectiveness of the recommendation are demonstrated by extensive simulation studies and real data examples.

E0228: **Learning high dimensional multi-response linear models with quantum optimization**
Presenter: **Yuan Ke**, University of Georgia, United States

A hybrid quantum computing algorithm is used to study linear regression models for high-dimensional multi-response data. An intuitively appealing estimation method is proposed based on identifying the linearly independent columns in the coefficient matrix. The approach relaxes the low-rank constraint in the existing literature and allows the rank to diverge with dimensions. A novel quantum optimization algorithm selects the linearly independent columns significantly faster than classical methods implemented on electronic computers are proved. The proposed estimation procedure enjoys desirable theoretical properties. Intensive numerical experiments are also conducted to demonstrate the finite sample performance of the proposed method, and a comparison is made with some popular competitors. The results show that this method outperforms all alternative techniques under various circumstances.

E0376: **fastkqr: A fast algorithm for kernel quantile regression**
Presenter: **Boxiang Wang**, University of Iowa, United States
Co-authors: Qian Tang, Yuwen Gu

Quantile regression is a powerful tool for robust and heterogeneous learning and has been used in a wide spectrum of applied areas. However, its computational cost can be prohibitively high in contemporary large-scale applications due to the nonsmoothness of the quantile loss. A new algorithm named fastkqr is introduced, which provides a significant advance toward computing quantile regression in reproducing kernel Hilbert spaces. The crux of fastkqr is a novel finite smoothing algorithm, which magically gives the exact quantile regression solutions rather than approximations. An interpretability issue of quantile regression, which involves fitted curves crossing at multiple quantile levels in finite samples, is also addressed by presenting a new algorithm for fitting the non-crossing kernel quantile regression by imposing penalizations on the kernel coefficients. Extensive simulations and real applications are used to demonstrate that fastkqr achieves the same accuracy as the state-of-the-art algorithms but can be orders of magnitude faster.

E0392: **Statistical analysis for a penalized EM algorithm in high dimensional mixture linear regression model**
Presenter: **Ning Wang**, Beijing Normal University, China

The expectation-maximization (EM) algorithm and its variants are widely used in statistics. In high-dimensional mixture linear regression, the model is assumed to be a finite mixture of linear regression forms, and the number of predictors is much larger than the sample size. The standard EM algorithm, which attempts to find the maximum likelihood estimator, becomes infeasible. A penalized EM algorithm is devised, and its statistical properties are studied. Existing theoretical results of regularized EM algorithms often rely on dividing the sample into many independent batches and employing a fresh batch of samples in each iteration of the algorithm. The algorithm and theoretical analysis do not require sample-splitting. The method and theory are also extended to multivariate response cases. The proposed methods also have encouraging performances in numerical studies.

E0978: **Optimal false discovery control of minimax estimators**
Presenter: **Qifan Song**, Purdue University, United States

Two major research tasks lie at the heart of high-dimensional data analysis: accurate parameter estimation and correct support recovery. The existing literature mostly aims for either the best parameter estimation or the best model selection result. However, little has been done to understand the potential interaction between the estimation precision and the selection behaviour. The minimax result shows that an estimator's type I error control performance directly links with its L_2 estimation error rate and reveals a trade-off phenomenon between the rate of convergence and the false discovery control: to achieve better accuracy, one risks yielding more false discoveries. In particular, the false discovery control behaviour of rate optimal and rate suboptimal estimators under different sparsity regimes are characterized, and a rigid dichotomy is discovered between these two estimators under near-linear and linear sparsity settings. In addition, a rigorous explanation is provided for the incompatibility phenomenon between selection consistency and rate minimaxity, which has been frequently observed in the high-dimensional literature.

EO083 Room 708 NOVEL STATISTICAL APPROACHES TO BRAIN SIGNALS AND IMAGES Chair: Hernando Ombao

E0967: **Gaussian random fields on networks and metric graphs**
Presenter: **David Bolin**, King Abdullah University of Science and Technology, Saudi Arabia

A new class of Gaussian processes on compact metric graphs such as street or river networks are defined. The proposed models, the Whittle-Matern fields, are defined via a fractional stochastic partial differential equation on the compact metric graph and are a natural extension of Gaussian fields with Matern covariance functions on Euclidean domains to the non-Euclidean metric graph setting. The existence of the processes, as well as their sample path regularity properties, are derived. The model class, in particular, contains differentiable Gaussian processes. To the best of our knowledge, this is the first construction of a valid differentiable Gaussian field on general compact metric graphs. Then it is focused on a model subclass which is shown to contain processes with Markov properties. In this case, it is shown how to evaluate finite-dimensional distributions of the process exactly and computationally efficiently. This facilitates using the proposed models for statistical inference without the need for any approximations. Finally, some of the main statistical properties of the model class, such as consistency of maximum likelihood estimators of model parameters and asymptotic optimality properties of linear prediction based on the model with misspecified parameters, are derived. The usage of the model class is illustrated through an application to traffic data.

E0975: **Low-rank and sparse decomposition for brain functional connectivity in naturalistic fMRI data**
Presenter: **Chee-Ming Ting**, Monash University, Malaysia, Malaysia
Co-authors: Jeremy Skipper, Fuad Noman, Steven Small, Hernando Ombao

A novel, data-driven approach is presented, which is based on low-rank plus sparse (L+S) decomposition to isolate stimulus-driven dynamic changes in fMRI brain functional connectivity (FC) from the background noise by exploiting shared network structure among subjects receiving the same naturalistic stimuli. The time-resolved multi-subject FC matrices are modelled as a sum of a low-rank component of correlated FC patterns across subjects and a sparse component of subject-specific, idiosyncratic background activities. To recover the shared low-rank subspace, a fused version of principal component pursuit (PCP) is introduced by adding a fusion-type penalty on the differences between the rows of the low-rank matrix. The method improves the detection of stimulus-induced group-level homogeneity in the FC profile while capturing inter-subject variability. An efficient algorithm is developed via a linearized alternating direction method of multipliers to solve the fused PCP. Simulations show accurate recovery by the fused-PCP even when a large fraction of FC edges is severely corrupted. When applied to natural fMRI data, the method reveals FC changes that were time-locked to auditory processing during movie watching, with the dynamic engagement of sensorimotor systems for speech-in-noise. It also provides a better mapping to auditory content in the movie than ISC.

E1106: **Bayesian image analysis in Fourier space for neuroimaging**
Presenter: **John Kornak**, University of California, San Francisco, United States
Co-authors: Karl Young, Eric Friedman, Konstantinos Bakas, Hernando Ombao

For more than 30 years now, Bayesian image analysis has been a leading approach to image reconstruction and enhancement. The idea of the approach is to balance a priori expectations of image characteristics (the prior) with a model for the image degradation process (the likelihood). The conventional Bayesian modelling approach as defined in image space, implements priors that describe inter-dependence between spatial locations on the image lattice (commonly through Markov random field, MRF, models) and can therefore be difficult to model and compute. Bayesian image analysis in Fourier space (BIFS) provides for an alternate approach that can generate a wide range of models, including ones with similar properties to conventional models but with a reduced computational burden; the originally complex high-dimensional estimation problem in image space can be similarly modelled as a series of (trivially parallelizable) independent one-dimensional problems in Fourier space. Development of different prior models in Fourier space will be examined, and it is illustrated with neuroimaging applications of BIFS applied to 1) longitudinal structural MRI for evaluating brain tumour evolution and 2) diagnosis of frontotemporal dementia based on perfusion-weighted MRI and PET.

E1262: **Filtrated common functional principal component analysis of multi-group functional data**
Presenter: **Shuhao Jiao**, City University of Hong Kong, Hong Kong
Co-authors: Ron Frostig, Hernando Ombao

Local field potentials (LFPs) are signals that measure electrical activities in localized cortical regions and are collected from multiple tetrodes implanted across a patch on the surface of the cortex. In many cases, multi-tetrode LFP trajectories contain both global variation patterns and idiosyncratic variation patterns, and such structure is very informative to the data mechanism. Therefore, one goal is to develop an efficient algorithm that is able to capture and quantify both global and idiosyncratic features. The novel filtrated common functional principal components (filt-fPCA) method is developed, a novel forest-structured fPCA for multi-group functional data. A major advantage of the proposed filt-fPCA method is its ability to extract the common components in a flexible multi-resolution manner. The proposed approach is highly data-driven, and no prior knowledge of ground-truth data structure is needed, making it suitable for analyzing complex multi-group functional data. In addition, the filt-fPCA method is able to produce parsimonious, interpretable, and efficient functional reconstruction (low reconstruction error) for multi-group functional data with orthonormal basis functions. The proposed filt-fPCA method is employed to study the impact of a shock (induced stroke) on the synchrony structure of rats' brains.

E0942: **Covariate-guided mixture of multivariate time series experts for interpretable analysis of fNIRS data**
Presenter: **Robert Krafty**, Emory University, United States
Co-authors: Haoyi Fu, Lu Tang, Ori Rosen, Alison Hipwell, Theodore Huppert

Similar to other measures of brain function, functional near-infrared spectroscopy (fNIRS) data take the form of heterogeneous multivariate time series signals. A novel group-based method for analyzing fNIRS simultaneously clusters subject-level data into potentially interpretable phenotypes is discussed while evaluating associations with clinical and demographic variables. The method models subject-level fNIRS data through a mixture of nonparametric time components where mixing weights depend on time-independent exogenous variables and account for heterogeneity among subjects. The proposed method is motivated by and illustrated through data analysis from a study of infant emotional reactivity and recovery from stress.

EO228 Room 709 STATISTICAL ML IN CYBERSECURITY Chair: Lekha Patel

E0229: **Unsupervised attack pattern detection in honeypot data using Bayesian topic modelling**
Presenter: **Francesco Sanna Passino**, Imperial College London, United Kingdom
Co-authors: Anastasia Mantziou, Daniyar Ghani, Philip Thiede, Ross Bevington, Nick Heard

Cyber systems are under near-constant threat from intrusion attempts. Attack types vary, but each attempt typically has a specific underlying intent, and the perpetrators are typically groups of individuals with similar objectives. Clustering attacks appearing to share a common goal is very valuable to threat-hunting experts. The purpose is to explore topic models for clustering terminal session commands collected from honeypots, which are special network hosts designed to entice malicious attackers. The main practical implications of clustering the sessions are two-fold: finding similar groups of attacks and identifying outliers. A range of statistical topic models is considered and adapted to the structures of command-line syntax. In particular, concepts of primary and secondary topics, and then session-level and command-level topics, are introduced into the models to improve interpretability. The proposed methods are further extended in a Bayesian nonparametric fashion to allow unboundedness in the vocabulary size and the number of latent intents. The methods are shown to discover an unusual MIRAI variant that attempts to take over existing cryptocurrency coin-mining infrastructure, which is not detected by traditional topic-modelling approaches.

E0595: **General-purpose unsupervised cyber anomaly detection via non-negative tensor factorization**
Presenter: **Maksim Eren**, Los Alamos National Laboratory, United States
Co-authors: Juston Moore, Erik Skau, Elisabeth Moore, Manish Bhattarai, Gopinath Chennupati, Boian Alexandrov

Distinguishing malicious anomalous activities from unusual but benign activities is a fundamental challenge for cyber defenders. Prior studies have shown that statistical user behaviour analysis yields accurate detections by learning behaviour profiles from observed user activity. These unsupervised models can generalize to unseen types of attacks by detecting deviations from normal behaviour without the knowledge of specific attack signatures. However, approaches proposed to date based on probabilistic matrix factorization are limited by the information conveyed in a two-dimensional space. On the other hand, non-negative tensor factorization is a powerful unsupervised machine learning method that naturally models multi-dimensional data, capturing complex and multi-faceted details of behaviour profiles, allowing us to improve the sensitivity and specificity for anomaly detection tasks. The new unsupervised statistical anomaly detection methodology matches or surpasses state-of-the-art supervised learning baselines across several challenging and diverse cyber application areas, including detecting compromised user credentials, botnets, spam e-mails, and fraudulent credit card transactions. Our methodology is based on our SmartTensors AI project (winner of R&D100 2021), a platform for accurate tensor decomposition algorithms that can scale to extra-large datasets.

E0634: **Statistical properties of compression analytics**
Presenter: **Kurtis Shuler**, Sandia National Laboratories, United States
Co-authors: Alexander Foss, Christina Ting, Travis Bauer, Richard Field

Compression analytics (CA) uses file compression algorithms to perform many predictive and inferential tasks typically associated with statistics and machine learning, such as clustering, anomaly detection, and classification. Unlike more traditional approaches, CA does not require explicitly defined covariates or engineered features but can be applied to any set of arbitrary bitstreams. CAs great flexibility allows it to be rapidly prototyped, tested, and deployed across a wide range of problems and domains, but its black-box nature has hindered connections to existing statistical theory. For lossless or near-lossless compression, this disconnect can be bridged by relating a bitstreams compression ratio to an explicit or implied model likelihood, enabling a wide variety of existing statistical theories and techniques to be applied to CA. As examples of how these connections can be employed, this relationship is exploited to show how existing model selection techniques such as AIC and BIC can be utilized in CA and develop a novel EM-like CA clustering algorithm. Finally, the efficacy of these algorithms is demonstrated by applying these CA techniques to both real and simulated datasets.

E0706: **Multi-attribute utility elicitation for real-time network anomaly detection**
Presenter: **Fletcher Christensen**, University of New Mexico, United States
Co-authors: Erin Schwertner-Watson, Lekha Patel, Gabriel Huerta, Douglas McGeehan

Identifying cyber attacks as they happen requires effective models for anomaly detection, but computational constraints may limit the space of features available in real-time. Multiple-criteria decision-making (MCDM) is a method for making decisions in situations where the utility of a decision is based on multiple criteria: for example, the accuracy of the model fit and computational burden. Statistical model selection traditionally identifies the 'best' model in a candidate class, where model utility is operationalized as out-of-sample model fit and information criteria act as an effective proxy for out-of-sample model fit. The aim is to examine how AIC, a common information criterion, can be modified according to multi-attribute utility theory (MAUT) to select models according to both out-of-sample model fit and computational burden. Some simple classes of multi-attribute utilities and elicitation techniques for verifying that these utilities agree with expressed preferences and with the axioms of the von Neumann-Morgenstern utility theorem are discussed.

E0888: **Local estimation and testing of latent network curvature**
Presenter: **Tyler McCormick**, University of Washington, United States
Co-authors: Steven Wilkins-Reeves

Network data, commonly used throughout the physical, social, and biological sciences, consists of nodes (individuals) and the edges (interactions) between them. One way to represent network data's complex, high-dimensional structure is to embed the graph into a low-dimensional geometric space. The curvature of this space, in particular, provides insights into the structure in the graph, such as the propensity to form triangles or present tree-like structures. An estimating function is derived for curvature based on triangle side lengths and the midpoints between sides where the only input is a distance matrix and also establishes asymptotic normality. Next, a novel latent distance matrix estimator has been introduced for networks and an efficient algorithm to compute the estimate via solving iterative quadratic programs. This method is applied to the Los Alamos National Laboratory Unified Network and Host dataset, and it is shown how curvature estimates can be used to detect a red-team attack faster than naive methods, as well as discover non-constant latent curvature in co-authorship networks in physics.

EC323 Room 02 DATA MINING AND EDUCATION Chair: Brenda Betancourt

E1025: **Mining and modeling the attendance of performing arts activities**
Presenter: **Tsung-Chi Cheng**, National Chengchi University, Taiwan

This research analyzes data from a survey which focuses on the related issues regarding the attendance of performing arts events in three cities located in northern Taiwan. Performing arts in Taiwan are classified into four categories: music, dance, contemporary drama, and traditional theatre. The aim is to mine patterns about how the audiences consume various types and combinations of performing arts activities through the association rule as well as identify determinants related to the event's attendance. Both the association mining approach and appropriate statistical modelling analytics are applied to achieve the purposes.

E1203: **Growth parameter estimation: A multiple mark and recapture analysis**
Presenter: **Chuan Foo**, Sultan Idris Education University, Malaysia

Various approaches have been employed in previous studies on lobster growth modelling, such as mark-recapture and tag-recapture methods. However, several limitations and challenges are associated with these methods, including low recapture rates and data quality issues, which hinder accurate estimates of lobster growth and development. Growth parameters for multiple recaptures of wild slipper lobster, Scyllarides latus, are estimated using the generalised estimating equation (GEE) and improved Fabens method. GEE incorporates a biologically realistic framework to efficiently describe the correlation between two consecutive moults, including the hidden variables about data derived from multiple recaptures. The improved Fabens method considers individual variability in growth where different individuals have different asymptotic lengths. Simulation results indicate that the Fabens method is slightly better than the GEE approach as it reduces the biases in the estimates. The results suggest that lobsters can exhibit significant variability in growth rates over their lifetimes.

E0625: **Analyzing successive knowledge sharing in virtual communities**
Presenter: **Shiu-Wan Hung**, National Central University, Taiwan

Knowledge-sharing behaviour in 28 virtual communities over three consecutive years. Hierarchical linear modelling was used to test 184 valid respondents. Results showed that longer interaction time among members led to increased norms, trust, and reciprocity, which in turn promoted knowledge-sharing behaviour. Community norms encouraged knowledge sharing in the external environment, while community trust and reciprocity were found to effectively enhance knowledge-receiving and providing behaviour.

E0344: **Determinants of early career teachers retention intention: Evidence from a multilevel model**
Presenter: **Mike Smet**, KU Leuven, Belgium

Numerous studies in different countries find increasing teacher turnover rates, leading to shortages and negatively impacting the quality of education. Especially the attrition rates of early career teachers are high and are a major cause for concern. The method examines early career teacher retention intention using data from the Teaching and Learning International Survey (TALIS) 2018, which OECD conducted. The empirical analysis includes 6,500 early career teachers nested in more than 2,300 schools, further nested in 20 regions. A multilevel or hierarchical linear model (HLM) is estimated to investigate the relationship between various individual and school-specific factors and the intention to stay among teachers.

Results show that several individual-level and school-level predictors (e.g. age, diploma, intrinsic or extrinsic motivation to become a teacher, type of contract, job satisfaction, satisfaction with salary, perception of workload, school composition, school climate, school size, school location, etc.) are significant predictors of early career teacher retention intention. The findings have important policy implications for reducing teacher attrition rates. Several significant predictors can be influenced by policy decisions, both at the school level or at a more aggregate level.

E0921: **The use of video in the teaching of statistics after the Covid19 pandemic: A case study with tourism students**
Presenter: **Sonia Pais**, Polytechnic of Leiria, Portugal

After the lockdown imposed by the Covid19 pandemic, millions of students and teachers faced new challenges as they progressively returned to their face-to-face school activities. Despite all the constraints caused by the pandemic, this is an opportunity to move towards an innovative and transformative educational model for the future, which must avoid sending us back to the negative aspects of past educational practices. During the lockdown, with emergency remote teaching (ERT) being implemented, the teacher involved in this study faced the need to adapt activities and didactic materials to be used exclusively at distance with the aid of technology. This included the promotion of asynchronous tasks/activities, which could be developed autonomously by students. Technologies have multiple benefits in the field of Education, such as improving student performance. Information and Communication Technologies are tools that may be used to innovate the way statistics is taught, and they may facilitate students' learning. Being aware of these benefits, part of the activities and teaching materials prepared for the ERT, in the face-to-face teaching of her classes after lockdown, more specifically with the presentation of videos for content development, are adapted and introduced. The participants in the study are undergraduate students from a Portuguese higher education institution enrolled in the curricular unit of Statistical Analysis.

EC320 Room 701 FORECASTING IN APPLICATIONS Chair: Rogemar Mamon

E1213: **Forecasting Philippine economic growth using mixed frequency data: MIDAS versus MF-DLFM**
Presenter: **Rutcher Lacaza**, University of the Philippines, Philippines

Assessing the economic impact of COVID-19 in most developing countries like the Philippines has been hampered by the delayed publication of official statistics, such as GDP. Traditional forecasting models for economic growth rely on aggregating economic or financial indicators observed at higher frequencies than quarterly GDP growth, which can lead to a loss of useful forward-looking information and less accurate forecasts. To address this issue, mixed-frequency models have been developed, such as the Mixed Data Sampling (MIDAS) regression technique and the Mixed-frequency Dynamic Latent Factor Model (MF-DLFM), to incorporate high-frequency data into prediction models. The aim is to compare the performance of MIDAS and MF-DLFM in forecasting quarterly GDP in the Philippines using monthly and weekly data from 2000 to 2023. The results indicate that both models outperform traditional models that use only quarterly data. However, the MF-DLFM provides slightly more accurate forecasts than the MIDAS model. The findings demonstrate the usefulness of mixed frequency models in providing timely and accurate information to policymakers, enabling informed decisions, especially during the COVID-19 pandemic.

E0226: **Modeling and forecasting of sales in fuel retail market: The factors that boost a fuel station's sales**
Presenter: **Marco Costa**, University of Aveiro, Portugal
Co-authors: Daniel Magueta, Stephanie Espadilha

Sales of a fuel company with a dense fuel station network were analyzed to identify and characterize potential variables with predictive capacity for sales of new fuel stations. The database consists of a set of context variables with predictive potential for sales of fuel stations and monthly sales in terms of fuel volume. The research methodology focused on multivariate statistical methods combining cluster analysis, regression models and forecasting models. The fuel station context variables tend to characterize the socio-economic conditions, such as population density variable, others related to the similar existing supply of both the company itself and the competing companies, and others related to geographical location and accessibility. The data analysis allowed us to identify clusters in the time series of sales, indicating that the investigation of factors must be segmented. Homogeneous groups of fuel stations were identified through a hierarchical agglomerative clustering procedure. For each of the groups identified, multiple linear regression models were adjusted considering the fuel sales in the first years of operation of the stations as dependent variables. It is possible to conclude that the average daily traffic is the variable with a higher predicted capacity for most of the groups of fuel stations analyzed.

E0235: **Forecasting tyre sales: Competitive models to improve inventory in a small business**
Presenter: **Magda Monteiro**, University of Aveiro, Portugal
Co-authors: Diana Neves, Maria Jose Felicio

Demand forecasting is one of the key aspects of operations management, in particular sales forecasting, as it plays a key role in the retailer's resource planning that impacts consumer satisfaction. In the current global competitive business environment, where customer service and timely delivery are critical factors, it is necessary to obtain accurate forecasts that reduce uncertainty in customer demand. Different competitive models are applied using simple models, such as exponential smoothing and also more complex models, to forecast tyre sales in a small business in order to choose the most suitable for presenting inventory plans for tyre sales. Tyre types were grouped according to their sales into half-yearly, quarterly and monthly to evaluate which forecasting models are best suited, within these groups, to forecast sales to use in defining the inventory plan. To compare the accuracy and efficiency of the competitive models, several measurements such as Root Mean Square Error, the final inventory level average and shortage percentages are used.

E0285: **Time series forecasting approaches to retail sales in EU Countries: Portugal and her major trading partners**
Presenter: **A Manuela Goncalves**, University of Minho, Portugal
Co-authors: Susana Lima, Marco Costa

In the area of economics, particularly in the retail segment, sales forecasting supports most of the strategic planning decisions of any retail business. It must be as accurate as possible to ensure corporate profitability. The main goal is to evaluate forecasting methods' accuracy in the area of time series modelling applied to retail segment data. A method is proposed to compare the ARMA model's accuracy and their extensions, the classical decomposition time series associated with multiple linear regression models, and the exponential smoothing methods (Holt-Winters). These methods are chosen because of their ability to model trends and seasonal fluctuations present in economic data, particularly in retail sales data. The data available on the Eurostat platform correspond to monthly indexes of EU Countries' retail trade turnover. According to PORDATA, Portugal's major trading partners (import and export of goods and services) are Germany, Spain, France, Italy, the Netherlands, and the UK. The results suggest that multiple linear regression models are not the most appropriate to forecast retail sales indexes, while the SARIMA models are identified as the most accurate ones. Holt-Winters models are also a viable alternative, although they are not considered the most appropriate.

E0986: **Forecasting of expenditures from foreign tourists traveling to Thailand**
Presenter: **Thidaporn Supapakorn**, Kasetsart University, Thailand
Co-authors: Sukanya Intarapak, Witchanee Vuthipongse

The objectives are to find a suitable forecasting model and forecasting period of the expenditure from foreign tourists traveling to Thailand. The data is gathered from January 2011 to December 2019 and is divided into two sets. The first set is the data from January 2011 to December

2018 for the modelling by the method of Box-Jenkins, Artificial Neural Network and combined forecasting of Box-Jenkins and Artificial Neural Network. The second is the monthly data for 2019 for comparing the performance of the forecasting models via the criteria of the lowest mean absolute percentage error (MAPE) and the root mean square error (RMSE). The results show that the combined model is the most accurate with the short-term (3 months) forecasting period, with the lowest MAPE and RMSE of 3.09% and 6,212.43 million baht, respectively.

Wednesday 02.08.2023 10:50 - 12:30 Parallel Session G – EcoSta2023

EV253 Room 701 TIME SERIES AND SPATIAL STATISTICS (VIRTUAL) Chair: Kun Chen

E0282: **Multiple testing of local extrema for detecting change points under nonstationary Gaussian noise**
Presenter: **Dan Cheng**, Arizona State University, United States

A new approach to detect change points based on differential smoothing and multiple testing is presented for data sequences modelled as piecewise constant functions plus nonstationary Gaussian noise. The method detects change points as significant local maxima and minima after smoothing and differentiating the observed sequence. The algorithm, combined with the Benjamini Hochberg procedure for thresholding p values, provides asymptotic strong control of the False Discovery Rate (FDR) and power consistency as the frequency of observations and the size of the jumps get large. Simulations show that FDR levels are maintained in non-asymptotic conditions and guide the choice of smoothing bandwidth.

E1197: **Vector error correction models with stationary and nonstationary variables**
Presenter: **Pu Chen**, Melbourne Institute of Technology, Australia

Vector error correction models (VECM) have become a standard tool in empirical economics for analysing nonstationary time series data because they combine two key concepts in economics: equilibrium and dynamic adjustment, in one single model. The current standard VECM procedure is restricted to time series data with the same degree of integration, i.e. all I(1) variables. However, time series data with different degrees of integration are common in empirical studies, necessitating the simultaneous handling of I(1) and I(0) time series. The standard VECM is extended to accommodate mixed I(1) and I(0) variables. The mixed VECM conditions are derived, and as a result, a test and estimation of the mixed VECM are presented.

E0324: **Statistical inference with stochastic gradient methods under ϕ-mixing data**
Presenter: **Ruiqi Liu**, Texas Tech University, United States
Co-authors: Xi Chen, Zuofeng Shang

Stochastic gradient descent (SGD) is a scalable and memory-efficient optimization algorithm for large datasets and stream data, which has drawn a great deal of attention and popularity. The applications of SGD-based estimators to statistical inference, such as interval estimation, have also achieved great success. However, most of the related works are based on i.i.d. observations or Markov chains. When the observations come from a mixing time series, how to conduct valid statistical inference remains unexplored. The general correlation among observations imposes a challenge on interval estimation. Most existing methods may ignore this correlation and lead to invalid confidence intervals. A mini-batch SGD estimator is proposed for statistical inference when the data is ϕ-mixing. The confidence intervals are constructed using an associated mini-batch bootstrap SGD procedure. Using the "independent block" trick from [**?**], it is shown that the proposed estimator is asymptotically normal, and the bootstrap procedure can effectively approximate its limiting distribution. The proposed method is memory-efficient and easy to implement in practice. Simulation studies on synthetic data and an application to a real-world dataset confirm the theory.

E1263: **Inference for spatial autoregressive models using stochastic gradient descent**
Presenter: **Ji Meng Loh**, New Jersey Institute of Technology, United States
Co-authors: Gan Luan

Using stochastic gradient descent (SGD), the procedure is considered to fit spatial auto-regressive models to lattice data, incorporating a recently developed perturbation method to obtain standard errors in addition to model parameter estimates. The SGD update equations are derived, and the results of a simulation study will be presented to examine the empirical coverage of confidence intervals constructed using the perturbation procedure.

EI002 Room 102 MATHEMATICAL STATISTICS AND TIME SERIES ANALYSIS Chair: Yan Liu

E0161: **A stochastic maximal inequality and its applications**
Presenter: **Yoichi Nishiyama**, Waseda University, Japan

It is well known that maximal inequalities play a key role in the fields such as weak convergence of random fields and high-dimensional statistics. Some approaches based on Bernstein's inequality for martingales have been successfully taken to obtain maximal inequalities for martingales. A new inequality, which may be called a stochastic maximal inequality, is presented. The inequality is a bound for maxima of a finite number of martingales by the sum of a predictable increasing process and a martingale starting from zero. It may be regarded as an inequality version of the Doob-Meyer decomposition. As its applications, some analogues of Doob's inequality and Lenglart's inequality to finite-dimensional martingales are obtained. Some applications to statistics will also be discussed.

E0963: **Statistical Inference for Glaucoma Detection**
Presenter: **Ngai Hang Chan**, City University of Hong Kong, Hong Kong

Some of the statistical techniques used in the analysis and detection of one of the most commonly encountered eye diseases, Glaucoma, will be reviewed. By means of some of the recent artificial intelligence algorithms, ophthalmologists are employing modern machine learning technologies such as CNN and its variates to assist in the early detection of eye symptoms in Glaucoma studies. Some of the related issues will be addressed, and the potential of statistical ideas in these studies will be discussed. Some examples will be given.

E1023: **Learning from similar linear representations: Adaptivity, minimaxity, and robustness**
Presenter: **Yang Feng**, NYU, United States

Representation multi-task learning (MTL) and transfer learning (TL) have achieved tremendous success in practice. However, the theoretical understanding of these methods is still lacking. Most existing theoretical works focus on cases where all tasks share the same representation and claim that MTL and TL almost always improve performance. However, assuming all tasks share the same representation as the number of tasks grows is unrealistic. Also, this does not always match empirical findings, which suggest that a shared representation may not necessarily improve single-task or target-only learning performance. The aim is to understand how to learn from tasks with similar but not exactly the same linear representations while dealing with outlier tasks. Two algorithms are proposed that are adaptive to the similarity structure and robust to outlier tasks under both MTL and TL settings. The algorithms outperform single-task or target-only learning when representations across tasks are sufficiently similar and the fraction of outlier tasks is small. Furthermore, they always perform no worse than single-task learning or target-only learning, even when the representations are dissimilar. Information-theoretic lower bounds are provided to show that the algorithms are nearly minimaxed optimal in a large regime.

EO124 Room 02 NEW DEVELOPMENTS IN MICROBIOME RESEARCH Chair: Julia Fukuyama

E0249: **Quantifying major sources of technical variability in microbiome sequencing lab protocols**
Presenter: **Ekaterina Smirnova**, Virginia Commonwealth University, United States

A significant issue with the horizontal harmonization of previously collected microbiome data is the large variation in the data processing through non-standardized methods. Microbiome study is a complex process that starts with sample collection and storage, followed by transportation to a DNA extraction and sequencing lab, running a bioinformatics pipeline to identify microbial taxa, and finally, statistical analysis of a summarized taxonomic table. This leads to large differences in microbial taxa even for the replicate samples. The previously collected Microbiome Quality Control Project (MBQC) data is utilized that process identical stool and artificial communities aliquots by 16 sample handling laboratories to quantify the primary sources of technical variability on downstream statistical analysis. All sequences are re-processed with the identical bioinformatics pipeline and rank the differences in alpha and beta diversity by sample handling protocols. The ultimate goal of this analysis is to inform the harmonization of previously collected microbiome studies and identify the major differences in microbiome sequencing protocols that must be accounted for in the pulled studies analysis.

E0439: **A flexible zero-inflated Poisson-Gamma model with application to microbiome sequence count data**
Presenter: **Tianying Wang**, Colorado State University, United States
Co-authors: Roulan Jiang, Xiang Zhan

In microbiome studies, it is of interest to use a sample from a population of microbes, such as the gut microbiota community, to estimate the population proportion of these taxa. However, due to biases introduced in sampling and preprocessing steps, these observed taxa abundances may not reflect true taxa abundance patterns in the ecosystem. Repeated measures, including longitudinal study designs, may be potential solutions to mitigate the discrepancy between observed abundances and true underlying abundances. Yet, widely observed zero-inflation and over-dispersion issues can distort downstream statistical analyses aiming to associate taxa abundances with covariates of interest. A Zero-Inflated Poisson Gamma (ZIPG) model framework is proposed to address these challenges above. From a perspective of measurement errors, the discrepancy between observations and truths is accommodated by decomposing the mean parameter in Poisson regression into a true abundance level and a multiplicative measurement of sampling variability is provided from the microbial ecosystem. A flexible ZIPG model framework by connecting both the mean abundance and the variability of abundances to different covariates and building valid statistical inference procedures for both parameter estimation and hypothesis testing. The proposed ZIPG method provides significant insights through comprehensive simulation studies and real data applications.

E0458: **Testing microbiome associations with survival times**
Presenter: **Yijuan Hu**, Emory University, United States

Finding microbiome associations with possibly censored survival times is an important problem, especially as specific taxa could serve as biomarkers for disease prognosis or as targets for therapeutic interventions. The existing methods are restricted to testing associations at the community level and do not provide results at the individual taxon level. An ad hoc approach testing each taxon with a survival outcome using the Cox proportional hazard model may not perform well in the microbiome setting with sparse count data and small sample sizes. The linear decomposition model (LDM) has been previously developed for testing continuous or discrete outcomes that unify community-level and taxon-level tests into one framework. The LDM is extended to test survival outcomes. The use of the Martingale residuals or the deviance residuals obtained is proposed from the Cox model as continuous covariates in the LDM. Further tests that combine the results of analyzing each set of residuals separately are constructed. Using simulated data, it is shown that the LDM-based tests preserved the false discovery rate for testing individual taxa and had good sensitivity. An analysis of data on the association of the gut microbiome and the time to acute graft-versus-host disease revealed several dozen associated taxa that would not have been achievable by any community-level test, as well as improved community-level tests by the LDM over existing methods.

E1137: **Context-aware dimensionality reduction of microbial ecosystem dynamics**
Presenter: **Liat Shenhav**, NYU, United States

Complex microbial ecosystems play an important role across many domains of life, from the female reproductive tract, through the oceans, to the plant rhizosphere. The study of these ecosystems offers great opportunities for biological discovery due to the ease of their measurement, the ability to perturb them, and their rapidly evolving nature. These same properties, however, make it difficult to extract robust and reproducible patterns from these high-dimensional and multi-scale environments. To address this, a context-aware dimensionality reduction method named Joint Compositional Tensor Factorization (Joint CTF) was developed that incorporates information from the same host across time, space and information layers (e.g., microbiome, metabolome, metatranscriptome). Joint CTF identifies robust patterns in longitudinal multi-omics data, allowing for the detection of ecosystem-wide changes associated with specific phenotypes that are reproducible across datasets. This model, designed to identify robust spatiotemporal patterns, would help us better understand the nature of the microbiome from the time of its formation and throughout life.

EO019 Room 03 NEW ADVANCES IN STATISTICAL LEARNING FOR IMAGE, PROCESS AND CLINICAL DATA Chair: Tiejun Tong

E0301: **Understanding differential item functioning using process data**
Presenter: **Ling Chen**, Columbia University, United States
Co-authors: Jingchen Liu

Differential item functioning (DIF) is an important concept in testing fairness. It occurs when items function differently among different subgroups. Previous research on DIF has mainly focused on statistical detection, yet understanding why DIF occurs remains a challenge. Process data obtained from respondents interacting with a computer-based assessment item provides a unique opportunity to understand DIF as it contains rich information about the progress and strategies towards problem-solving for each respondent. Using features extracted from process data, a variable that alleviates the DIF effect is constructed, which helps in detecting behavioural patterns that could lead to DIF and thus provides a deeper understanding of the underlying mechanism of DIF.

E0582: **Estimating the reciprocal of a binomial proportion**
Presenter: **Jiajin Wei**, Hong Kong Baptist University, Hong Kong
Co-authors: Ping He, Tiejun Tong

The binomial proportion is a classic parameter with many applications and has been extensively studied in the literature. By contrast, the reciprocal of the binomial or inverse binomial proportion is often overlooked, even though it also plays an important role in various fields. To estimate the inverse binomial proportion, the maximum likelihood method fails to yield a valid estimate when no successful event exists in the Bernoulli trials. To overcome this zero-event problem, several methods have been introduced in the previous literature. Yet to the best of our knowledge, there is little work on a theoretical comparison of the existing point estimators. Also, there is little work on the interval estimation for the inverse binomial proportion. To fill the gap, first, some commonly used point estimators for the inverse binomial proportion are reviewed, and then a new estimator is developed that aims to eliminate the estimation bias. Moreover, four different methods are applied to construct the confidence intervals (CIs), namely the Wald, score, arctangent and beta prime CIs, and further, their respective statistical properties are studied. Numerical studies are conducted to evaluate the finite sample performance of the proposed estimators, followed by a recent meta-analysis on the prevalence of heart failure among COVID-19 patients with mortality to demonstrate their usefulness in practice.

E0937: **Introduction to differencing method and the application in testing no effect**
Presenter: **Zhijian Li**, Beijing Normal University- Hong Kong Baptist University United International College, China

The idea of differencing to remove nonparametric effects provides a simple way to analyze the estimation and inference for nonparametric or semiparametric regression models. The difference-based methods are popular in practice and have many related applications. A novel difference-based method is introduced for testing the hypothesis of no relationship between the dependent and independent variables. Test statistics for nonparametric regression with Gaussian and non-Gaussian random errors are constructed. Further, it is demonstrated that these tests can detect local alternatives that converge to the null hypothesis at a rate close to n to the power -1/2. Simulation results show that the new tests are more powerful than existing methods, especially when the sample size is small.

E1076: **Scalable statistical inference in non-parametric least squares**
Presenter: **Meimei Liu**, Virginia Tech, United States
Co-authors: Yun Yang, Zuofeng Shang

Stochastic approximation (SA), such as stochastic gradient descent (SGD), is a powerful and scalable algorithm for solving stochastic optimization problems in large-scale and streaming data settings. An inferential framework for stochastic approximation (SA) in nonparametric least squares problems within a reproducing kernel Hilbert space (RKHS) is developed. An online multiplier bootstrap method is proposed for local inference through pointwise confidence intervals and global inference with simultaneous confidence bands, thus advancing SA-based estimation in non-parametric regression models. The main contributions include the development of a unified framework to derive the non-asymptotic behaviour of the infinite-dimensional stochastic gradient descent (SGD) estimate under the supremum norm, demonstrating the consistency of the multiplier bootstrap method in nonparametric settings. Further, the proposed method is applied to neuroimage data for statistical inference.

EO308 Room 04 OPTIMAL EXPERIMENTAL DESIGNS: RECENT ADVANCES AND APPLICATIONS Chair: Saumen Mandal

E1079: **Complex innovative design pilot program and a potential proposal**
Presenter: **Li Wang**, AbbVie, United States

The complex, innovative design (CID) pilot program from FDA provided an excellent opportunity for industry statisticians to think out of the box and to push the boundary on trial designs for late-stage studies. A potential Bayesian CID in SLE is proposed to combine Bayesian dose-ranging, master protocol, sequential group design and short-term/long-term biomarkers together to speed up clinical development and increase the probability of success for the platform. Design parameters, including the maximum sample size, are calibrated to obtain good frequentist properties such as type I error and power. A hypothetical trial of three agents for systemic lupus erythematosus is used to illustrate the concept, and extensive simulations show that the proposed design compares favourably to several conventional platform designs.

E0640: **PUMP: Estimating power when adjusting for multiple outcomes in multi-level experiments**
Presenter: **Kristen Hunter**, UNSW, Australia
Co-authors: Kristin Porter, Luke Miratrix

For randomized controlled trials (RCTs) with a single intervention's impact being measured on multiple outcomes, researchers often apply a multiple testing procedure (MTP) (such as Bonferroni or Benjamini-Hochberg) to adjust p-values. Such an adjustment reduces the likelihood of spurious findings but also changes the statistical power, sometimes substantially, which reduces the probability of detecting effects when they do exist. However, this consideration is frequently ignored in typical power analyses, as existing tools do not easily accommodate the use of MTPs. The PUMP R package is introduced as a tool for analysts to estimate statistical power, minimum detectable effect size (MDES), and sample size requirements for multi-level RCTs with multiple outcomes. One of PUMP's main innovations is accommodating multiple outcomes: power estimates from PUMP properly account for the adjustment in p-values from applying an MTP. Also, PUMP allows researchers to consider a variety of definitions of power in order to choose the most appropriate types of power for the goals of their study. The package supports a variety of commonly-used frequentist multi-level RCT designs and linear mixed effects models. In addition to the main functionality of estimating power, MDES, and sample size requirements, the package allows the user to easily explore the sensitivity of these quantities to changes in underlying assumptions.

E0812: **Optimal designs for matching adjusted indirect comparison**
Presenter: **Saumen Mandal**, University of Manitoba, Canada
Co-authors: Xiang Zheng

As part of a clinical trial, a new treatment is compared with a competitor treatment to determine its effect on the patient. Ideally, the new treatment can be directly compared with the competitor treatment in randomized controlled trials. However, it is difficult to directly compare due to various factors, such as time, price, regulation, and patents. A matching-adjusted indirect comparison method leverages all available data by adjusting average patient characteristics in trials with individual patient data (IPD) to match those reported in the aggregate trials data. As IPD is used to match the pre-defined baseline characteristics, optimal design theory is used, and this is converted into a constrained optimization problem. A Lagrangian method is used to determine the optimal design subject to satisfying the constraints of baseline characteristics. The new methodology is quite flexible and can be applied to different types of constraints. The methodology can be applied to situations where there is a lack of direct comparison. It will also reduce the time and cost of running experiments.

E0823: **An algorithm for searching optimal variance component estimators in linear mixed models**
Presenter: **Subir Ghosh**, University of California, United States

An algorithm is developed to derive strictly positive unbiased estimators with minimum variance in a well-defined class. The key idea is to take the method-of-moments estimator, given by a quadratic form of a symmetric matrix A, as a starting point and modify it using the class of square non-singular regularization matrices Q while preserving unbiasedness in addition to ensuring positivity. Different subclasses of structured Q are possible for convenience instead of all possible Q matrices. A search algorithm then finds a local or global optimal matrix A depending on Q and the corresponding optimal variance component estimator by minimizing the variance, a function of the unknown variance components and kurtosis parameters. The proposed method further allows finding matrices A leading to quadratic forms closely approximating the corresponding numerical values of the likelihood-based estimates. In addition, the dependence of variance functions is investigated on unknown model parameters. Using two illustrative examples, Examples I and II, the report also illustrates the use of the matrix Q and the determination of the kurtosis parameters in the search for the optimal variance component estimators by keeping the bias zero and the variance trim for the bias-variance trade-offs.

EO028 Room Virtual R01 NEW CHALLENGES FOR COMPLEX AND LARGE-SCALE DATA Chair: Yichuan Zhao

E0279: **An efficient variance estimator for cross-validation under partition-sampling**
Presenter: **Qing Wang**, Wellesley College, United States
Co-authors: Xizhen Cai

The problem of variance estimation of cross-validation is concerned. It considers an unbiased cross-validation risk estimator in the form of a general U-statistic. The proposal is an efficient variance estimator under a half-sampling design, where the estimator's bias can be expressed explicitly. Furthermore, one can approximate the bias by a two-layer Monte Carlo method so that a bias-corrected variance estimator can be obtained. In the simulation study and real data examples, the performance of the proposed variance estimator, in comparison to the commonly used bootstrap and jackknife methods, is evaluated in the context of model selection. The numerical results suggest that the proposal yields identical or similar conclusions for model selection compared to its counterparts, and it is much more efficient to calculate than its competitors. The generalization of the methodology to other partition-sampling scenarios is also discussed.

E0591: **Two tools for preliminary data analysis: Entropic plots for tail identification and heatmap for variable clustering**
Presenter: **Jialin Zhang**, Mississippi State University, United States

First, a non-parametric method with entropic plots is introduced to identify the thickness of tails for the underlying discrete distributions. Based on the sample, the method identifies the thickness of the tail and compares it with Four benchmark tails: 1) power decaying tails, 2) sub-exponential decaying tails, 3) near-exponential decaying tails, and 4) exponential decaying tails. The first three benchmarks are considered thick tails. When an underlying discrete distribution is identified as a thick-tailed distribution, the method can provide a point estimation and an interval estimation for the parameters to assist further parametric analysis. Second, a non-parametric method with Generalized Shannons Entropy (GSE) is introduced to identify the similarities among ordered discrete distributions. The underlying ordered discrete distributions are not restricted to a common sample space. The method is theoretically supported by a characterization theorem with a set of GSEs. Based on the GSEs, a heatmap can be constructed to provide a simple and intuitive visualization to compare the ordered probability distributions for all the discrete random variables to help obtain a preliminary understanding of the variables.

E0950: **A non-smoothing framework for inference on functional means**
Presenter: **Hsin-wen Chang**, Academia Sinica, Taiwan

A nonparametric inference framework that applies to occupation time curves derived from wearable device data is introduced. Motivated by the right continuity of these curves, a non-smoothing approach is developed that involves weaker conditions than existing conditions imposed when using smoothing to estimate functional means under a fixed dense design. Notably, the procedure allows discontinuities in the functional covariances while accommodating the discretization of the observed trajectories. Under this non-smoothing framework, an empirical likelihood method is devised to construct confidence bands for the functional means. The method utilizes the known optimality of empirical likelihood. It also respects range and monotonicity constraints on occupation time curves. A simulation study shows that the proposed procedures outperform competing functional data procedures.

E0997: **Jackknife empirical likelihood inference for the accelerated failure time model**
Presenter: **Yichuan Zhao**, Georgia State University, United States

The accelerated failure time (AFT) model is a useful semi-parametric model under right censoring, which is an alternative to the commonly used proportional hazards model. Making statistical inferences for the AFT model has attracted considerable attention. However, it is difficult to compute the estimators of regression parameters due to the lack of smoothness for rank-based estimating equations. Brown and Wang (2007) used an induced smoothing approach, which smooths the estimating functions to obtain point and variance estimators. A more computationally efficient method called jackknife empirical likelihood (JEL) is proposed to make inferences for the accelerated failure time model without computing the limiting variance. Results from extensive simulation suggest that the JEL method outperforms the traditional normal approximation method in most cases. Subsequently, two real data sets are analyzed to illustrate the proposed method.

EO140 Room Virtual R02 RECENT DEVELOPMENTS FOR MODELING HIGH-DIMENSIONAL AND COMPLEX DATA Chair: Wenbo Wu

E1281: **On partial envelop approach for modeling spatial-temporally dependent data**
Presenter: **Reisa Widjaja**, University of Texas at San Antonio, United States

Modeling multivariate spatial-temporally dependent data is a challenging task due to the dimensionality of the features and the complex spatial-temporal associations among the data. A parsimonious approach is used by proposing a spatial-temporal partial envelop model to achieve efficient estimations in modelling the spatial-temporal data. The model is extended to a group-wise spatial-temporal partial envelop model to adjust the heterogeneity existing at different locations. It is provided with both theoretical justifications and conducts thorough empirical simulations to demonstrate the effectiveness of the proposed method. The proposed model is also applied to analyzing the crowdsourcing weather data collected from personal weather stations in the United States.

E1287: **On sufficient variable screening using log odds ratio filter**
Presenter: **Wenbo Wu**, University of Texas at San Antonio, United States

For ultrahigh-dimensional data, variable screening is an important step to reduce the scale of the problem, hence, improving the estimation accuracy and efficiency. A new dependence measure which is called the log odds ratio statistic to be used under the sufficient variable screening framework. The sufficient variable screening approach ensures the sufficiency of the selected input features in modelling the regression function and is an enhancement of existing marginal screening methods. In addition, an ensemble variable is the proposed screening approach to combine the proposed fused log odds ratio filter with the fused Kolmogorov filter to achieve supreme performance by taking advantage of both filters. The sure screening properties of the fused log odds ratio filter for both marginal variable screening and sufficient variable screening are established. Extensive simulations and a real data analysis are provided to demonstrate the usefulness of the proposed log odds ratio filter and the sufficient variable screening procedure.

E1289: **Reproducible learning for accelerated failure time models via deep knockoffs**
Presenter: **Daoji Li**, California State University Fullerton, United States

Selecting truly relevant variables contributing to the response is a fundamental problem in many scientific fields. One of the major challenges in variable selection is effectively controlling the false discovery rate (FDR). Most existing variable selection procedures in survival analysis neglect the FDR control. Such a gap is filled, and a new and flexible variable selection method with guaranteed FDR control is proposed for accelerated failure time models. The proposed method combines the strengths of deep knockoffs and the weighted M-estimation procedure and enjoys the FDR control for arbitrarily high dimensions with finite samples. More importantly, the proposed method does not require prior knowledge about the joint distribution of covariates. Extensive simulation studies confirm the proposed method's generality, effectiveness, and power. Finally, the proposed method is used to analyze primary biliary cirrhosis data to demonstrate its practical utility.

E1298: **Distributed instrumental variable analysis in three UK studies**
Presenter: **Yanchun Bao**, University of Essex, United Kingdom
Co-authors: Hongsheng Dai, Wei Liang

In observational data analysis, the instrumental variable analysis is a popular approach to obtain the causal effects of a risk factor (exposure) X on an outcome (response) Y. An instrument variable G is correlated with the exposure X but not correlated to the confounder U or directly correlated to the response Y. This is also called Mendelian randomization when genetic variants are used as the instrument variable to examine the causal effect of a modifiable exposure on a particular disease in an observational study. Merging results from different data sources is challenging for such studies based on instrument variables. One reason is that the data privacy barriers do not allow data from different studies to be transferred and stored for centralized analysis. Distributed analyses have to be implemented. However, the heterogeneity of the instrument variable G indifferent studies makes it very challenging to combine results from different data sources into a final conclusion using a naive meta-analysis approach or divide-and-conquer approach. A novel distributed analysis with instrumental genetic variables is developed to overcome the heterogeneity of the instrument variables in different studies. Simulation and data applied in three UK studies have been used to illustrate the proposed method.

EO185 Room 201 RECENT ADVANCES IN CAUSAL INFERENCE AND MISSING DATA Chair: BaoLuo Sun

E0168: **Nonparametric mediation analysis: Beyond the mean**
Presenter: **Yen-Tsung Huang**, Academia Sinica, Taiwan

Mediation analyses estimate the effect of exposure on an outcome mediated by a mediator. Many methods have been developed to conduct mediation analyses, and most of them focus only on the mean outcome. The biological effect of a cancer mutation may have affected the variation of a downstream outcome, such as gene expression and patient survival, without even changing the mean outcome. To this end, the effect on the cumulative distribution function (cdf) of an outcome is characterized. Consequently, one can easily summarize the impact of an outcome on any specific moment, with the effect on the mean outcome as a special case. A nonparametric estimator is proposed based on kernel estimators for the cdf of the mediator given the exposure and that of the outcome given the exposure and mediator. The uniform consistency and weak convergence of the proposed estimators are established. Extensive simulation studies were conducted to evaluate the performance of finite samples. These methods are applied to two studies: one investigates the influence of childhood socioeconomic adversity on adult adiposity via DNA methylation of the FASN (fatty acid synthase) gene, and another investigates how IDH1 (isocitrate dehydrogenase 1) mutations in glioma patients affect EGFR (epidermal growth factor receptor) expression by altering its DNA methylation.

E0762: **A novel penalized inverse-variance weighted estimator for Mendelian randomization with applications to COVID-19 outcomes**
Presenter: **Zhonghua Liu**, Columbia University, United States

Mendelian randomization utilizes genetic variants as instrumental variables (IVs) to estimate the causal effect of an exposure variable on an outcome of interest, even in the presence of unmeasured confounders. However, the popular inverse-variance weighted (IVW) estimator could be biased in the presence of weak IVs, a common challenge in MR studies. A novel penalized inverse-variance weighted (pIVW) estimator, which adjusts the original IVW estimator, is developed to account for the weak IV issue by using a penalization approach to prevent the denominator of the pIVW estimator from being close to zero. Moreover, the variance estimation of the pIVW estimator is adjusted to account for the presence of balanced horizontal pleiotropy. It is shown that the recently proposed debiased IVW (dIVW) estimator is a special case of the proposed pIVW estimator. Further, it is proved that the pIVW estimator has a smaller bias and variance than the dIVW estimator under some regularity conditions. Extensive simulation studies and real data analysis are also conducted to demonstrate the performance of the proposed pIVW estimator.

E0976: **Double negative control inference in test-negative design studies of vaccine effectiveness**
Presenter: **Xu Shi**, University of Michigan, United States

The test-negative design (TND) has become a standard approach to evaluating vaccine effectiveness. Despite TND's potential to reduce unobserved differences in healthcare-seeking behaviour (HSB) between vaccinated and unvaccinated subjects, it remains subject to various potential biases. First, residual confounding bias may remain due to unobserved HSB, occupation as a healthcare worker, or previous infection history. Second, because selection into the TND sample is a common consequence of infection and HSB, collider stratification bias may exist when conditioning the analysis on testing, which further induces confounding by latent HSB. Third, the generalizability of the results to the general population is not guaranteed. A novel approach is presented to identify and estimate vaccine effectiveness in the general population by carefully leveraging a pair of negative control exposure and outcome variables to account for potential hidden bias in TND studies. The proposed method is illustrated with extensive simulation and an application to COVID-19 vaccine effectiveness using data from the University of Michigan Health System.

E1128: **Sparse causal mediation analysis with unmeasured mediator-outcome confounding**
Presenter: **Wei Li**, Renmin University of China, China

Causal mediation analysis aims to investigate how an intermediary factor called a mediator regulates the causal effect of a treatment on an outcome. With the increasing availability of measurements on a large number of potential mediators in various disciplines, methods for conducting mediation analysis with many or even high-dimensional mediators have been proposed. However, they often assume there is no unmeasured confounding between mediators and the outcome. Such confounding is allowed, and an approach is provided to address both identification and mediator selection problems under the structural equation modelling framework. The identification strategy involves constructing a pseudo proxy variable for unmeasured confounding based on a latent factor model for multiple mediators. Using this proxy variable, a partially penalized procedure is proposed to select important mediators with nonzero effects on the outcome. The resultant estimates are consistent, and the estimates of nonzero parameters are asymptotically normal. Simulation studies show the advantageous performance of the proposed procedure over other existing methods. Finally, this approach is applied to genomic data, and gene expressions that may actively mediate the effect of a genetic variant on mouse obesity are identified.

EO202 Room 203 STATISTICAL ADVANCES IN GENETIC EPIDEMIOLOGY Chair: Debashree Ray

E0251: **Integrative analysis of multiple microbiome studies**
Presenter: **Ni Zhao**, Johns Hopkins University, United States

Recent studies have highlighted the importance of human microbiota in our health and diseases. However, in many research areas, individual microbiome studies often offer inconsistent results due to the limited sample sizes and the heterogeneity in study populations and experimental procedures. Integrative analysis of multiple microbiome datasets is necessary. However, statistical methods that incorporate multiple microbiome datasets and account for the study heterogeneity are unavailable in the literature. Two recent developments are discussed from our lab that aims at the integrative analysis of multiple microbiome datasets, one for the analysis of alpha diversity and one for the analysis of beta diversities. These approaches are applied to data from the HIV-reanalysis consortium, a collective effort that obtained all publicly available data on the gut microbiome and HIV in December 2017, and a coherent association of gut microbiome is obtained with HIV infection and with MSM status (i.e. men who have sex with men).

E0372: **Meta-analysis in family-based study of disease subtypes**
Presenter: **Debashree Ray**, Johns Hopkins University, United States

Family-based designs often examine genetic determinants of child health outcomes and rare diseases (e.g., case-parent trio design consisting of an affected child and both parents). When investigating similarities and differences in the genetic basis of subtypes of such diseases, investigators have typically used meta-analysis techniques. However, a meta-analysis in this context tests the global null hypothesis that a genetic marker does not affect any disease. This is not exactly the null hypothesis to test when the goal is to identify a common genetic basis. A new statistical approach is discussed for detecting the genetic overlap of two diseases or disease subtypes by considering a composite null hypothesis that a genetic marker is associated with none or only one of the traits. A mixture distribution is used for the null distribution of the test statistic that allows for fractions of millions of genetic markers to be associated with none or only one of the traits. An asymptotic approximation of the null distribution IS useD that avoids estimating nuisance parameters related to mixture proportions and variance components. Our method requires only summary-level data as used in the meta-analysis, has well-calibrated type I error at stringent levels used in genetic studies, and can achieve major power gain over alternative methods typically used in the literature. Finally, an application is shown to the case-parent trio study of orofacial cleft subtypes.

E0446: **Knockoff-based statistics for the identification of putative causal loci in genetic studies**
Presenter: **Iuliana Ionita-Laza**, Columbia University, United States

Knockoff-based methods are becoming increasingly popular due to their enhanced power for locus discovery and their ability to prioritize putative causal variants in a genome-wide analysis. However, generating knockoffs is computationally and memory expensive, and applying this methodology to genetics is nontrivial. Scalable knockoff-based methods for biobank-sized data for population-based designs and related extensions to family-based designs are discussed. Applications are shown in several large-scale genomic studies, including the UK biobank data.

E0887: **Polygenic risk score methods to improve disease screening**
Presenter: **John Witte**, Stanford University, United States

Polygenic risk scores (PRS) combine information across large numbers of genetic variants to give an individualized genetic susceptibility profile that may be useful for disease prediction. PRS can also be leveraged to improve biomarker and screening accuracy. For example, genetically adjusting prostate-specific antigen (PSA) with a PRS for this biomarker improves prostate cancer screening. Specifically, diagnostic decisions based on PSA values adjusted using a PRS would have avoided 31% of negative prostate biopsies but also resulted in 12% fewer biopsies in prostate cancer cases, mostly in patients with Gleason score <7 tumours. Important outstanding questions surrounding the use of PRS will be considered, including the most appropriate PRS modelling approach, especially to maximize PRS transferability across populations. PRS developed in populations of European genetic ancestries may have poor predictive performance in individuals of other ancestries; thus, their use can potentially increase health disparities. Recent advances are considered in PRS modelling, including Bayesian shrinkage of the risk score weights and SuperLearner ensemble methods across multiple PRS approaches. Such innovative developments can help to generate PRS that benefit diverse populations.

EO164 **Room 503** **RECENT ADVANCES IN DESIGN AND ANALYSIS OF COMPLEX EXPERIMENTS** **Chair: Luke Keele**

E0326: **Supervised stratified subsampling: An approach to big data predictive analytics**
Presenter: **Ming-Chung Chang**, Academia Sinica, Taiwan

Predictive analytics encompasses the use of statistical models for prediction. Its power, however, is hindered by the rising amounts of data in recent years. Owing to advanced technology, big data are ubiquitous across disciplines. Such data richness may yield difficulties in predictive analytics either in terms of time cost or numerical stability. A new subsampling approach is introduced to overcome this difficulty for regression problems. The proposed method integrates a nonparametric regression technique and stratified sampling, referred to as supervised stratified subsampling. Theoretical properties are developed to justify this method. Numerical studies show that the proposed method yields good predictions and is against model misspecification.

E0720: **A projection space-filling criterion and related optimality results**
Presenter: **Chenlu Shi**, Colorado State University, United States
Co-authors: Hongquan Xu

Computer experiments call for space-filling designs. Recently, a minimum aberration type space-filling criterion was proposed to rank and assess a family of space-filling designs, including Latin hypercubes and strong orthogonal arrays. It aims to capture a design's space-filling properties when projected onto subregions of various sizes. The dimension aside from the sizes of subregions by proposing first an expanded space-filling hierarchy principle and then a projection space-filling criterion as per the new principle are also considered. When projected onto subregions of the specific size, the proposed criterion ranks designs via sequentially maximizing the space-filling properties on equally-sized subregions in lower dimensions to higher dimensions, while the minimum aberration type space-filling criterion compares designs by maximizing the aggregate space-filling properties on multidimensional subregions of the same size. Further, the construction of the optimal space-filling designs is considered under the proposed criterion. Although many algorithms have been proposed for generating space-filling designs, it is well-known that they often deteriorate rapidly in performance for large designs. In the present paper, some theoretical optimality results and characterize several classes of strong orthogonal arrays of strength three that are the most space-filling are developed.

E1089: **Thompson sampling with discrete support**
Presenter: **Wei Zheng**, University of Tennessee, United States

Thompson sampling is a popular algorithm for multi-armed bandit problems, but its Bayesian posterior update can be computationally expensive for complex reward distributions. Recently, prior discretization has been proposed to address this issue. A new prior discretization method is proposed that guarantees the same regret rate without requiring the unreasonable assumption that the true value of the parameter is one of the discrete points. Additionally, a modified posterior update approach is introduced, improving the performance of discrete prior Thompson sampling. It is proven that the accumulated regret has a $O(log(T))$ convergence rate with high probability. In addition, numerical experiments are conducted to validate the theoretical analysis and demonstrate that the proposed algorithm outperforms both the standard discrete prior method and the Laplace approximation approach for the continuous prior.

E1107: **A maximin Φ_p-efficient design for multivariate generalized linear models**
Presenter: **Yiou Li**, DePaul university, United States

Experimental designs for a generalized linear model (GLM) often depend on the model's specification, including the link function, the predictors, and unknown parameters, such as the regression coefficients. To deal with the uncertainties of these model specifications, it is important to construct optimal designs with high efficiency under such uncertainties. Existing methods, such as Bayesian experimental designs, often use prior distributions of model specifications to incorporate model uncertainties into the design criterion. Alternatively, one can obtain the design by optimizing the worst-case design efficiency with respect to the uncertainties of model specifications. A new Maximin Φ_p-Efficient (or Mm-Φ_p for short) design is proposed, aiming to maximize the minimum Phi_p-efficiency under model uncertainties. Based on the theoretical properties of the proposed criterion, an efficient algorithm with sound convergence properties is developed to construct the Mm-Φ_p design. The performance of the proposed Mm-Φ_p design is assessed through several numerical examples.

EO249 Room 506 STATISTICS ON MANIFOLDS Chair: Tomonari Sei

E0929: **Sinkhorn diffusion and Wasserstein mirror gradient flows**
Presenter: **Nabarun Deb**, University of British Columbia, Vancouver, Canada

The sequence of marginals obtained from iterations of the Sinkhorn or IPFP algorithm on joint densities converge is proved, under suitable time and parameter scaling and other assumptions, to be an absolutely continuous curve on the Wasserstein space. The limit is an example of the Wasserstein mirror gradient flow, a construction inspired by the Euclidean mirror gradient flows. In the case of Sinkhorn, the gradient is that of relative entropy. The parabolic Monge Ampere PDE provides an equivalent description of this flow, whose connection to Sinkhorn was noticed before by Berman. A Mckean-Vlasov SDE whose marginal distributions give the same flow is constructed; and can be viewed as the mirror analogue of the Langevin diffusion.

E0727: **Mixture of normalizing flows for spherical density estimation**
Presenter: **Tin Lok James Ng**, Trinity College Dublin, Ireland
Co-authors: Andrew Zammit Mangion

The use of normalizing flows to model complex probability distributions has attracted much research interest in the machine-learning community in recent years. Normalizing flows offer great flexibility in modelling probability distributions, only requiring the specification of a simple reference distribution and a series of bijective transformations. More recently, research interests have shifted to developing normalizing flows for probability distributions on spaces with more complex geometries, such as spheres. However, using a global normalizing flow to model complex probability distributions proved challenging in some applications. The aim is to extend this framework to a mixture model by using normalizing flows as mixture components for density estimation on the sphere.

E0799: **Theoretical properties of log-concave projections in CAT(0) orthant space**
Presenter: **Yuki Takazawa**, The University of Tokyo, Japan
Co-authors: Tomonari Sei

Orthant space is a space consisting of multiple nonnegative Euclidean orthants that are glued together on common faces. An important example of CAT(0) orthant spaces is the space of phylogenetic trees. One of the critical statistical challenges is the construction and estimation of probability distributions in these spaces. However, due to the complexity of the space, it is not simple to construct a parametric family of distributions. Shape-constrained density estimation is a method used to estimate distributions by imposing constraints on the shape of densities. One of the commonly used shape constraints in Euclidean spaces is log-concavity. Although the class of log-concave distributions is nonparametric, the estimation by maximum likelihood is possible. The generalization of this method to the space of phylogenetic trees has been proposed previously, and further generalization to CAT(0) orthant spaces is straightforward. This research investigates some properties of log-concave projections in CAT(0) orthant spaces. Log-concave projection refers to the log-concave density that minimizes the Kullback-Leibler divergence from a given probability measure. First, a sufficient condition is given for the existence of log-concave projections, and their uniqueness is shown. Then, by deriving a certain continuity property of a Kullback-Leibler type functional, conditions for the consistency of the maximum likelihood estimators to the log-concave projections are derived.

E1034: **Minimum information dependence modeling for mixed-domain data analysis**
Presenter: **Keisuke Yano**, The Institute of Statistical Mathematics, Japan
Co-authors: Tomonari Sei

A method of constructing a joint statistical model for mixed-domain data is proposed to analyze their dependence. Multivariate Gaussian and log-linear models are particular examples of the proposed model. The model is characterized by two orthogonal sets of parameters: the parameters of dependence and those of marginal distributions. The existence and uniqueness theorem is presented for the proposed model. To estimate the dependence parameter, Conditional inference is established and its consistency is shown. Also, the information-geometrical structure and the connection to the entropic optimal transport and the Schrodingerbridge problems are discussed. Finally, an application is illustrated to the earthquake data.

EO014 Room 603 STATISTICAL ANALYSIS IN CRIME, INSURANCE AND PRODUCTION Chair: Boris Choy

E0659: **The effect of parole supervision on recidivism in New South Wales, Australia**
Presenter: **Joanna Wang**, University of Technology Sydney, Australia, Australia

The aim is to estimate the causal impact of parole supervision on recidivism among offenders sentenced to short-term prison sentences. The method used was to compare recidivism rates between parolees and ex-inmates who were released from prison unconditionally. The variation in the sentencing severity of quasi-randomly assigned Local Court magistrates as an instrument of release on parole is used to measure the causal effect of parole supervision. The results showed that parolees were substantially less likely to re-offend than prisoners released unconditionally. The likelihood of re-conviction, committing a person, property or serious drug offence, and being re-imprisoned within 12 months of release was significantly reduced. These reductions in recidivism persisted 24 months after release from prison. Furthermore, the findings revealed statistically significant reductions in recidivism among parolees with Medium or above LSI-R scores and below Medium and among Aboriginal and non-Aboriginal parolees. In conclusion, it is found that parole supervision had a substantial and lasting impact on reducing recidivism among offenders sentenced to short-term prison sentences.

E0671: **Modelling COVID and crime in the US as hierarchical time series**
Presenter: **Thomas Fung**, Macquarie University, Australia
Co-authors: Joanna Wang

Crime time series data can often be naturally disaggregated by various attributes of interest, either by their crime type or geographical location. When modelling this type of data, the current recommended practice in crime science is to model each series at the most disaggregated level as it helps to identify more subtle changes. However, authorities and stakeholders are often only interested in the big picture, requiring researchers to either simply sum up the fitted value series or model the aggregated series independently. This leads to poor forecasting performance at the higher levels of aggregation in practice, as the most disaggregated series often have a high degree of volatility, while the most aggregated time series is usually smooth and less noisy. Intuition also requires the forecasts to add up in the same way as the data, but one can't guarantee that would be the case when series are modelled independently. The aim is to explain why a previous hierarchical and grouped time series method should be considered as the default technique for modelling this kind of data. US COVID and crime data will be used as an example.

E0690: **Loss reserving using geometric processes**
Presenter: **Shiying Gao**, The University of Sydney, Australia
Co-authors: Boris Choy, Junbin Gao

Forecasting the future liability of an insurance product is a risk management exercise in an insurance company. Under-estimation of loss reserves may result in insolvency. The claim data presented is modelled in a run-off triangle using a stochastic model. As observed in many real claim data, claim amounts show trend patterns across the accident years and development years. The novelty is to model the trends using the geometric process. The Bayesian approach is used in data analysis for statistical inference and forecast of loss reserves. The performance of the geometric process in modelling the trends shall be investigated.

E1245: **Stochastic frontier analysis with scale mixtures of normal distributions**
Presenter: **Boris Choy**, University of Sydney, Australia

Stochastic frontier analysis has been widely used as an econometric model to estimate the production frontier and to measure the inefficiency of a company. In a stochastic frontier model, it has two random error terms. The random error for inefficiency is one−sided while the random error for the noise is two−sided. Different combinations of distributions for these two random errors are considered, and their performance is compared in real examples.

EO048 **Room 604** RECENT DEVELOPMENTS ON PANEL DATA ANALYSIS	**Chair: Wendun Wang**

E0400: **Asymptotic properties of the synthetic control method**
Presenter: **Xiaomeng Zhang**, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, China
Co-authors: Wendun Wang, Xinyu Zhang

New insights into the asymptotic properties of the synthetic control method (SCM) are provided. It is shown that the synthetic control (SC) weight converges to a limiting weight that minimizes the mean squared prediction risk of the treatment-effect estimator when the number of pretreatment periods goes to infinity. The rate of convergence is also quantified. Observing the link between the SCM and model averaging, the asymptotic optimality of the SC estimator under imperfect pretreatment fit is further established in the sense that it achieves the lowest possible squared prediction error among all possible treatment effect estimators that are based on an average of control units, such as matching, inverse probability weighting and difference-in-differences. The asymptotic optimality holds regardless of whether the number of control units is fixed or divergent. Thus, the results provide justifications for the SCM in a wide range of applications. The theoretical results are verified via simulations.

E0461: **Panel quantile regression for extreme risk**
Presenter: **Xuan Leng**, Xiamen University, China

Panel quantile regression models play an essential role in finance, insurance, and risk management applications. However, a direct application of panel regression for the extreme conditional quantiles may suffer from significant estimation errors due to data sparsity on the far tail. A two-stage method is introduced to predict extreme conditional quantiles over cross-sections. First, use panel quantile regression at a selected intermediate level, then extrapolate the intermediate level to an extreme level with extreme value theory. The combination of panel quantile regression at an intermediate level and extreme value theory relies on a set of second-order conditions for heteroscedastic extremes. Also, a metric called Average Absolute Relative Error is proposed to evaluate the prediction performance of both intermediate and extreme conditional quantiles. Individual fixed effects in panel quantile regressions complicate the asymptotic analysis of the two-stage method and prediction metric. Compared to the direct use of panel quantile regression, the finite sample performance of the extreme conditional quantile prediction compared is demonstrated. Finally, the two-stage method is applied to the macroeconomic and housing price data, and strong evidence of housing bubbles and common economic factors is found.

E0721: **Debiased inference for nonlinear panel maximum-likelihood models with two-way fixed effects**
Presenter: **Yutao Sun**, Dongbei University of Finance and Economics, China
Co-authors: Xuan Leng

Panel data models often use fixed effects to account for unobserved heterogeneities. These fixed effects are typically incidental parameters, and their estimators converge slowly relative to the square root of the sample size. In the maximum likelihood context, this induces an asymptotic bias of the likelihood function. Test statistics derived from the asymptotically biased likelihood no longer follow their standard limiting distributions. This causes severe distortions in test sizes. A generic class of dynamic nonlinear models with two-way fixed effects is considered, and an analytical bias correction method for the likelihood function is proposed. It is formally shown that the likelihood ratio, the Lagrange multiplier, and the Wald test statistics derived from the corrected likelihood follow their standard asymptotic distributions. As a by-product, a bias-corrected estimator of the structural parameter can also be derived from the corrected likelihood function. The performance of the bias correction procedure is evaluated through simulations.

E1012: **Estimation of heterogeneous panel data models with an application to program evaluation**
Presenter: **Ke Miao**, Fudan University, China
Co-authors: Liangjun Su, Xun Lu

Panel data models with two-dimensional unobserved slope heterogeneity and interactive fixed effects are studied. A two-step approach is proposed to estimate the parameters in the model. In the first step, preliminary consistent estimators of the factors and factor loadings via a nuclear-norm-regularization (NNR) are obtained. In the second step, an iterative procedure is proposed to estimate the parameters of interest. The asymptotic properties of the estimators in each stage are established. The proposed model is applied to estimate heterogeneous treatment effects at both individual and aggregate levels. Monte Carlo simulations show that the proposed estimators perform remarkably well in finite samples compared to some existing methods, such as the synthetic control ones. The method is applied to study the effect of economic liberalization on economic growth and find a positive and significant aggregate average treatment effect on the treated (ATT).

EO108 **Room 605** NEW ADVANCES IN GAUSSIAN PROCESS MODELING AND COMPUTER EXPERIMENTS	**Chair: Chih-Li Sung**

E1086: **Dimension reduction for Gaussian process models via convex combination of kernels**
Presenter: **Lulu Kang**, Illinois Institute of Technology, United States

Many computer simulation models in engineering and scientific domains involve a high number of input variables, which can result in high computational costs and reduced prediction accuracy for the Gaussian process (GP) model. However, some simulation models may be influenced by only a small subset of the input variables, referred to as the active variables. Identifying these active variables can help researchers overcome the GP model's limitations and better understand the simulated system. To address this issue, a new approach for identifying the effective input dimensions of the GP model is proposed. Specifically, the covariance kernel function of the original GP model is approximated using a convex combination of kernels from lower-dimensional input dimensions. An iterative algorithm based on the Fedorov-Wynn algorithm from the optimal design literature was developed to determine the best approximation. The effect heredity principle while selecting the active input variables, which ensures that the subset of variables identified is sparse, is also incorporated. We demonstrate the effectiveness of the proposed method through several examples, showing that it outperforms some alternative approaches in correctly identifying the active input variables.

E0614: **Multi-fidelity Gaussian process modeling with boundary information**
Presenter: **Matthias Tan**, City University of Hong Kong, Hong Kong

Time-consuming bi-fidelity simulations with a high-fidelity (HF) simulator and a low-fidelity (LF) simulator, where the HF simulator contains a vector of inputs not shared with the LF simulator, called the augmented input, frequently arise in practice. For such simulations, it is frequently known a priori that when the augmented input converges to any value in a subset of the boundary of its domain, the HF simulator output converges to the LF simulator output. This is a form of boundary information, i.e., prior information on a simulator's output at the boundary of the domain of the simulator's inputs. The standard autoregressive Gaussian process (GP) emulator can be constructed to approximate and replace the HF and LF simulators to reduce simulation time. However, this emulator does not satisfy boundary information. A solution will be presented to the problem of constructing a bi-fidelity GP emulator that satisfies the form of boundary information just mentioned. The proposed emulator called the boundary-modified autoregressive GP (BMAGP) emulator, is shown to outperform the standard autoregressive GP emulator with some examples.

E1016: **Sobolev calibration of imperfect computer models**
Presenter: **Wenjia Wang**, HKUST (GZ), China

Calibration refers to the statistical estimation of unknown model parameters in computer experiments, such that computer experiments can match underlying physical systems. The aim is to develop a new calibration method for imperfect computer models, Sobolev calibration, which can rule out calibration parameters that generate overfitting calibrated functions. It is proved that the Sobolev calibration enjoys desired theoretical properties, including fast convergence rate, asymptotic normality and semiparametric efficiency. An interesting property is also demonstrated that the Sobolev calibration can bridge the gap between two influential methods: L_2 calibration and Kennedy and O'Hagan's calibration. In addition to exploring the deterministic physical experiments, the proposed method is theoretically justified that can transfer to the case when the physical process is indeed a Gaussian process, which follows the original idea of Kennedy and O'Hagan. Numerical simulations, as well as a real-world example, illustrate the competitive performance of the proposed method.

E1055: **Scalable and efficient computation of additive Gaussian processes with applications to Bayesian optimization**
Presenter: **Liang Ding**, Fudan University, China

Additive Gaussian Processes (GPs) are popular as priors in scalable high-dimensional Bayesian optimization. In Bayesian optimization, the next sampling point is determined by optimizing an adaptive acquisition function to the GP posterior. However, after sampling the n-th design point, updating the posterior requires $O(n^3)$ time. Compounding the value and gradient of the acquisition function requires $O(n^2)$, making it time-consuming, particularly for large n. While efficient algorithms exist for posterior updates of additive GPs, few studies have focused on the efficient computation of the acquisition function and its gradient. Since searching for the next sampling point requires iteratively acquiring the acquisition function and its gradient, it is more time-consuming than updating the posterior. Algorithms are proposed for posterior updates, hyperparameters learning, and computations of the acquisition function and its gradient of additive GPs with Matern covariances. The algorithms significantly reduce the time complexity of computing the acquisition function and its gradient from $O(n^2)$ to $O(\log n)$ for general learning rates and even to $O(1)$ for small learning rates, while achieving comparable computational efficiency to current best algorithms in updating the posterior and learning hyperparameters.

EO129 Room 606 ECONOMETRIC MODELING WITH TIME SERIES Chair: Tao Wang

E0237: **Model averaging factor-augmented quantile regressions with smooth structural change**
Presenter: **Siwei Wang**, Hunan University, China
Co-authors: Yundong Tu

Quantile regression is an effective tool in modelling data with heterogeneous conditional distribution. The time-varying coefficient quantile predictive regressions with factor-augmented predictors are considered to capture smooth structural changes and incorporate high-dimensional data information in prediction simultaneously. Theoretical results are established, including the uniform consistency and the asymptotic normality of the quantile estimators under misspecification. A novel time-varying jackknife model averaging method, which utilizes the local leave-one-out cross-validated weight, is developed to improve the forecast accuracy. The averaging estimator is asymptotically optimal in the sense of out-of-sample final prediction error. Numerical results demonstrate the superior performance of the averaging estimators.

E0391: **Kernel mode-based varying coefficient models with nonstationary regressors**
Presenter: **Tao Wang**, University of Victoria, Canada

Varying coefficient models are estimated on the basis of mode value using a kernel objective function, where the regressors are generated by multivariate unit root processes but can also be stationary. Such a kernel model-based estimation is demonstrated to be more robust and efficient than least squares estimation for data with outliers or heavy-tailed distributions, and it does not lose any efficiency when the data follow a normal distribution. A local linear approximation scheme is developed to estimate the varying coefficient function. Under mild regularity conditions, the asymptotic normality of the resultant estimators for both the unknown varying coefficient function and its derivative function is established. It is shown that the nonparametric estimator of the varying coefficient function with nonstationary regressors converges faster than the estimator with stationary regressors. In order to achieve estimation optimality in the sense of minimizing the asymptotic mean squared error, a kernel mode-based two-step estimation procedure is then suggested. The finite sample performance of the developed estimator is illustrated through three Monte Carlo simulations as well as a real data application on evaluating credit rationing in the United States credit market.

E1290: **Inference in linear models with structural changes and mixed identification strength**
Presenter: **Bertille Antoine**, Simon Fraser University, Canada
Co-authors: Otilia Boldea, Niccolo Zaccaria

Estimation and inference in a linear IV model in the presence of parameter instability are considered. When the reduced form is stable, but the structural form exhibits structural change, new GMM estimators are proposed, and it is proved that they are more efficient than the standard subsample GMM estimators, even in the presence of weaker identification patterns. For detecting change points in the structural form, two test statistics are proposed: when the reduced form is stable and when the reduced form exhibits structural change. The limiting distribution of these test statistics is derived, and it is shown that they have the correct asymptotic size and non-trivial power even under weaker identification patterns. The finite sample properties of the proposed estimators and testing procedures are illustrated in a series of Monte-Carlo experiments and in an application to the NKPC.

E1108: **Fluctuation-type monitoring test for explosive behavior**
Presenter: **Eiji Kurozumi**, Hitotsubashi University, Japan

A fluctuation-type monitoring test for a bubble is proposed. The initial value is dealt with by either OLS or quasi-difference demeaning. The asymptotic property of the test under mildly explosive and local alternatives is investigated. It is shown that the fluctuation-type test has an advantage over the existing methods when the bubble appears mid- to late in the monitoring period or the bubble period is relatively long, whereas the CUSUM monitoring scheme performs better in view of power for an early bubble in the monitoring period. This theoretical property is supported in finite samples by Monte Carlo simulations. As none of the existing tests uniformly outperforms the others, the union of rejections strategy by combining the two or three monitoring tests is also proposed, which is shown to work well in finite samples.

EO098 Room 702 DYNAMIC TOPOLOGICAL DATA ANALYSIS ON TIME SERIES DATA Chair: Moo K Chung

E0475: **Topological state-space estimation of dynamically changing functional human brain networks**
Presenter: **Moo K Chung**, University of Wisconsin-Madison, United States

A new data-driven topological approach is presented for estimating state spaces in dynamically changing functional brain networks of humans. The approach penalizes the topological distance between networks and clusters, dynamically changing brain networks into topologically distinct states. The method considers the temporal dimension of the data through the Wasserstein distance between networks. The method is shown to outperform the widely used k-means clustering often used in estimating the state space in brain networks. The method is applied to accurately determine the state spaces of dynamically changing functional brain networks. Subsequently, the question of if the overall topology of brain networks is a heritable feature using the twin study design is addressed.

E0557: **Spectral topological data analysis for EEG brain signals**
Presenter: **Hernando Ombao**, KAUST, Saudi Arabia
Co-authors: Anass El Yaagoubi Bourakna, Moo K Chung, Shuhao Jiao

Topological data analysis has become a powerful approach over the last twenty years, mainly because it captures the shape and geometry inherent in the data. Specifically, the use of persistence homology for analyzing functional brain connectivity has witnessed considerable success in the literature. It solves the problem of connectivity matrix thresholding at arbitrary levels by considering a filtration of the weighted network across all possible threshold values. Such approaches for analyzing the topological structure of functional brain connectivity rely on simple connectivity measures such as Pearson correlation. To overcome this limitation, a frequency-specific approach that leverages coherence is proposed to assess the brain's functional connectivity, leading to a novel topological summary, the spectral landscape, which is an extension of the persistence landscape. Using this novel approach to analyze the EEG brain connectivity of ADHD subjects, frequency-specific differences in the topology of brain connectivity between healthy controls and ADHD subjects are shed light.

E0904: **Topological data analysis of time-series data**
Presenter: **Jae-Hun Jung**, POSTECH, Korea, South

Time-series data analysis is found in various applications that deal with sequential data over a given interval of, e.g. time. Time-series data analysis is discussed based on topological data analysis (TDA). The commonly used TDA method for time-series data analysis utilizes embedding techniques such as sliding window embedding. With sliding window embedding, the given data points are translated into the point cloud in the embedding space, and the method of persistent homology is applied to the obtained point cloud. Some examples of time-series data analysis with TDA are first shown. Then, the recent work of exact and fast multi-parameter persistent homology (EMPH) theory will be introduced. The EMPH method is based on the Fourier transform of the data and the exact persistent barcodes. The EMPH is highly advantageous for time-series data analysis because its computational complexity is as low as O(Nlog N) and provides various topological inferences almost in no time.

E1283: **Topological clustering and inference on heat-diffusion estimates of persistence diagrams**
Presenter: **Yuan Wang**, University of South Carolina, United States
Co-authors: Jian Yin

Topological data analysis (TDA) has motivated the exploration of mesoscale features in brain signals and networks. Persistent homology is a key TDA algorithm for decoding and representing these features via persistence descriptors such as persistence diagram (PD), whose statistical significance is often revealed through permutation testing. But testing on PDs is challenging due to the heterogeneity of points in the diagrams that encode the birth and death times of features through a dynamic filtration of the subnetworks. The purpose is to showcase a topological clustering and transposition-based permutation testing framework based on heat-diffusion estimates of PDs to resolve computational bottlenecks of heavy permutations, with applications to the comparison of brain networks constructed from neuroimaging data.

EO103 Room 703 ADVANCES IN NETWORK DATA ANALYSIS Chair: Guodong Li

E0541: **An efficient tensor regression for high-dimensional data**
Presenter: **Guodong Li**, University of Hong Kong, Hong Kong

Most currently used tensor regression models for high-dimensional data are based on Tucker decomposition, which has good properties but loses its efficiency in compressing tensors very quickly as the order of tensors increases, say greater than four or five. However, for the simplest tensor autoregression in handling time series data, its coefficient tensor already has the order of six. A newly proposed tensor train (TT) decomposition is revised, and then it is applied to tensor regression such that a nice statistical interpretation can be obtained. The new tensor regression can well match the data with hierarchical structures, and it even can lead to a better interpretation of the data with factorial structures, which are supposed to be better fitted by models with Tucker decomposition. More importantly, the new tensor regression can be easily applied to the case with higher-order tensors since TT decomposition can compress the coefficient tensors much more efficiently. The methodology is also extended to tensor autoregression for time series data, and nonasymptotic properties are derived for the ordinary least squares estimations of both tensor regression and autoregression. A new algorithm is introduced to search for estimators, and its theoretical justification is also discussed. The theoretical and computational properties of the proposed methodology are verified by simulation studies, and the advantages over existing methods are illustrated by two real examples.

E0267: **Joint latent space models for ranking data and social network**
Presenter: **Jiaqi Gu**, Stanford University, United States
Co-authors: Philip Yu

Human interaction and communication has become one of the essential features of social life. Individuals' preferences may be influenced by those of their peers or friends in a social network. In the literature, individuals' rank-order preferences and their social networks are often modelled separately. A new joint probabilistic model is proposed for ranking data and social networks. With a latent space for all the individuals and items, the proposed model assumes that the social network and rankings of items are governed by the locations of individuals and items. Based on an efficient MCMC algorithm, a set of Bayesian inference approaches is developed for the proposed model, including procedures of model selection, criteria to evaluate model fitness and a test for conditional independence between individuals' rankings and their social network given their positions in the latent space. Simulation studies reveal the usefulness of the proposed methods for parameter estimation, model fitness evaluation, model selection and conditional independence testing. Finally, the model is applied to the CiaoDVD dataset, consisting of users' trust relations and implicit preferences on DVD categories.

E0951: **Preference matrix completion with multiple network views based on graph neural networks**
Presenter: **Philip Yu**, The Education University of Hong Kong, Hong Kong
Co-authors: Yipeng Zhuang, Chenlu Wang

In our digital age, people are connected to many networks. Their preferences (such as ratings and rankings) for all the items (such as movies and products) may not be complete and be affected by some of these networks. To predict the missing preferences, a novel graph neural network model will be proposed for preference matrix completion with multiple network views and side information. An attention mechanism was applied to measure the influences from multi-view networks. New objective functions are proposed to evaluate the preference prediction performance. Finally, the proposed model will be applied to some real-world movie recommendation datasets. Empirical results demonstrate that this model significantly improves preference prediction over other existing models.

EO088 Room 704 ECOLOGICAL STATISTICS MODELING Chair: Wen-Han Hwang

E0705: **Saturated pairwise interaction Gibbs point process as a joint species distribution model**
Presenter: **Yan Wang**, RMIT University, Australia
Co-authors: Ian Flint, Peter Vesk, Nick Golding, aihua xia

The saturated pairwise interaction Gibbs point process is introduced to model observed patterns in the spatial configuration of individuals of multiple species in nature. Its main strength lies in its ability to model attraction and repulsion within and between species over different scales. As such, it is particularly well-suited to studying associations in complex ecosystems. Based on the existing literature, an easy-to-implement fitting procedure and a technique to make inferences for the model parameters are provided. Different numerical experiments show the robustness of the model. Various ecological datasets are studied, demonstrating in each one that the model helps disentangle competing ecological effects on species distribution.

E0841: **Estimating population size: The importance of model and estimator choice**
Presenter: **Matthew Schofield**, University of Otago, New Zealand

The motivation comes from a mark-recapture distance sampling analysis. Unexpectedly large differences were found between Bayesian and frequentist estimates of abundance despite a moderately large number of observations (600). Further exploration revealed similar sensitivity to estimator choice when focusing on frequentist estimation. To understand these differences, abundance estimation from general mark-recapture models with three estimation strategies (maximum likelihood estimation, conditional maximum likelihood estimation, and Bayesian estimation) is considered for both binomial and Poisson capture-recapture models. It is found that assuming the data have a binomial or multinomial distribution introduces implicit and unnoticed assumptions that are not addressed when fitting with maximum likelihood estimation. This can have an important effect, particularly if the data arise from multiple populations. The results are compared to those of restricted maximum likelihood in linear mixed effects models.

E0853: **Nc-mixture occupancy models**
Presenter: **Wen-Han Hwang**, National Tsing Hua University, Taiwan

A class of occupancy models for detection/non-detection data is proposed to relax the closure assumption of N$-$mixture models. A community parameter c, ranging from 0 to 1, is introduced, which characterizes a certain portion of individuals being fixed across multiple visits. As a result, when c equals 1, the model reduces to the N$-$mixture model; this reduced model is shown to overestimate abundance when the closure assumption is not fully satisfied. Additionally, by including a zero-inflated component, the proposed model can bridge the standard occupancy model ($c = 0$) and the zero-inflated N$-$mixture model ($c = 1$). Then the behaviour of the estimators is studied for the two extreme models as c varies from 0 to 1. An interesting finding is that the zero-inflated N$-$mixture model can consistently estimate the zero-inflated probability (occupancy) as c approaches 0, but the bias can be positive, negative, or unbiased when $c > 0$ depending on other parameters. These results are also demonstrated through simulation studies and data analysis.

E1168: **Spatio-temporal joint modelling on moderate and extreme air pollution in Spain**
Presenter: **Chengxiu Ling**, Xián Jiaotong-Liverpool University, China

Very unhealthy air quality may cause numerous diseases. Extreme analysis and accurate predictions are in rising demand for exploring potential linked causes and for providing suggestions for environmental agencies. The aim is to model the spatial and temporal pattern of both moderate and extreme PM10 from 342 monitors throughout mainland Spain from 2017 to 2021. Bayesian hierarchical generalized extreme models of annual maxima PM10, including both fixed effects and spatiotemporal random effect with SPDE-AR(1) model, are proposed. The similar and different effects of interrelated factors are identified through a joint Bayesian model of annual mean and annual maxima PM10, which may bring the power of statistical inference of body data to the tail analysis with the implementation of the INLA algorithm. Under WAIC, DIC and other criteria, the best model is selected with good predictive ability. The findings are applied to identify the hot-spot regions with extremely poor air quality using excursion functions. It suggests that the community of Madrid and the northwestern boundary of Spain are likely to be exposed to severe air pollution, simultaneously exceeding the warning risk threshold. The joint model also provides evidence that precipitation, vapour pressure and population density influence comparably while altitude and temperature impact oppositely.

EO065 Room 705 THEORY AND METHODS FOR HIGH-DIMENSIONAL AND COMPLEX DATA Chair: Anuradha Roy

E0390: **Simple EM algorithm for Cauchy-type distributions**
Presenter: **Toshihiro Abe**, Hosei University, Japan
Co-authors: Masahiro Kuroda

An explicit expression of the EM algorithm is considered for the distribution defined by the Cauchy-type distributions. The main idea is developed by making use of the exponential distribution structure, from which a general estimator function is then derived. The explicit forms of the EM algorithm for the Cauchy, log-Cauchy and other Cauchy-type distributions are derived as examples. The simple updated formula of a skew-t distribution is also tackled, as well as its finite mixture and regression models.

E0397: **Quadratic classifiers for high-dimensional noisy data**
Presenter: **Aki Ishii**, Tokyo University of Science, Japan
Co-authors: Kazuyoshi Yata, Makoto Aoshima

One of the features of modern data is that the data dimension is extremely high. However, the sample size is relatively low. Such data is called HDLSS data. In HDLSS situations, new theories and methodologies are required to develop statistical inferences. High-dimensional classification is considered for noisy data such as genome data. It is noted that eigenvalues of high-dimensional noisy data grow very rapidly depending on the dimension. These eigenvalues obscure differences between populations. Two types of high-dimensional eigenvalue models exist the strongly spiked eigenvalue (SSE) model and the non-SSE (NSSE) model. High-dimensional classification under the SSE model is considered. New classifiers are given by using a data transformation technique. It is shown that our classifiers have preferable properties in theory. Finally, the classifiers are applied to real genome data sets.

E0528: **Supervised classification of high-dimensional data through functional data augmentation and random forest**
Presenter: **Fabrizio Maturo**, Universita Telem.universitas Mercatorum, Italy
Co-authors: Annamaria Porreca

With the advancement of technology, extensive amounts of data, such as sensor data, can now be collected over time, and analysing such data often requires supervised classification strategies. However, exploring high-dimensional data poses challenges, such as the curse of dimensionality and finding a balance between complexity and accuracy. To handle these challenges, researchers are investigating the combination of functional data analysis (FDA) and statistical learning. A supervised classification strategy that blends functional data augmentation and functional random forest techniques is presented. These methods are used to extract new features from high-dimensional data and produce augmented functional classifiers to enhance prediction accuracy. Novel interpretative rules in the functional domain are proposed based on separation rules introduced by exploiting derivatives information within the classification trees. Simulation studies and applications to real datasets demonstrate promising results, exceeding previously established accuracy records on online datasets.

E0575: **Weighted conditional network testing for multiple high-dimensional correlated data sets**
Presenter: **Takwon Kim**, Seoul National University, Korea, South
Co-authors: Inyoung Kim, Ki-Ahm Lee

Gaussian graphical models (GGMs) have been investigated to infer dependence (or network) structure among high-dimensional data by estimating a precision matrix. However, while many estimation methods for GGM have been developed, methods for testing the equality of two precision matrixes are still limited. Because testing the equality of the precision matrix depends on other given precision matrices, a weighted conditional network testing for considering other given precision matrices information is developed, and theoretical properties are also provided. None of the existing methods can be applied to test conditional differences when other networks are conditionally given and different. The advantage of the approach using a simulation study and genetic pathway analysis is demonstrated.

EO250 Room 708 NEW STATISTICAL METHODS IN NEUROIMAGING Chair: John Kornak

E0476: **Biclustering multivariate longitudinal data: Application to white matter recovery trajectories after sport-related mTBI**
Presenter: **Jaroslaw Harezlak**, Indiana University School of Public Health-Bloomington, United States
Co-authors: Luo Xiao

Biclustering is the task of simultaneously clustering the samples and features of a dataset. In doing so, subsets of samples that exhibit similar behaviours across subsets of features can be identified. Motivated by a longitudinal diffusion tensor imaging study of sport-related concussion (SRC), the problem of biclustering multivariate longitudinal data in which subjects and features are grouped simultaneously based on longitudinal patterns rather than a magnitude is presented. A penalized regression-based method is proposed for solving this problem by exploiting the heterogeneity in the longitudinal patterns within subjects and features. The performance of the proposed methods is evaluated via a simulation study, and an analysis of the motivating data set is performed. Subgroups of SRC cases that exhibit heterogeneous patterns of white-matter abnormalities are revealed.

E1115: **Linear effect of inter-scanner variability: Insights from paired cross-scanner T1-weighted images in elderly subjects**
Presenter: **Dana Tudorascu**, University of Pittsburgh, United States

Collecting structural MRI data across sites increases statistical power and enables the generalization of research outcomes; however, due to the variety of imaging acquisition, inter-scanner variability hinders the direct comparability of multi-scanner MRI data. Thus, many harmonization methods have been proposed to reduce inter-scanner variability in the image domain. Although proposed methods, especially incorporating deep learning techniques, have achieved promising performance, interpretability and understanding of inter-scanner variability were still limited. A small sample of eighteen cognitively normal participants is investigated, each scanned on four different 3T scanners, including GE, Philips, Siemens-Prisma and Siemens-Trio, during a short period of time (at most a few weeks apart). A statistical harmonization method, ComBat, was applied and extended to the image domain and investigated the linear effect of inter-scanner variability and image quality metrics. Furthermore, it is attempted to harmonize cross-scanner images by removing the estimated site effect. Besides estimating parametric maps of side effects, image quality metrics were calculated using MRIQC and similarity index to investigate the manifestation of scanner-related variation. Voxel-based morphometry using CAT12 was used to estimate cortical volumetric measures to compare the difference before and after the harmonization.

E1180: **Causality-based topological ranking of brain regions during epileptic seizure**
Presenter: **Anass El Yaagoubi Bourakna**, King Abdullah University of Science and Technology, Saudi Arabia
Co-authors: Moo K Chung, Hernando Ombao

The development of Topological Data Analysis over the last twenty years and its application to functional brain connectivity allowed neuroscientists to discover novel topological patterns in the organization of the brain. A novel TDA approach for oriented networks is proposed that intends to analyze the causality patterns in effective brain connectivity. Evidence is provided that the approach can detect causality patterns using simulated data. Furthermore, the application of this novel approach to epileptic seizure EEG data sheds light on the disruptive effect of seizures on the causal relationships between brain regions.

E1264: **Bayesian multi-object data integration in the study of primary progressive aphasia**
Presenter: **Aaron Scheffler**, University of California, San Francisco, United States
Co-authors: Rajarshi Guhaniyogi, Rene Gutierrez

Clinical researchers collect multiple images from separate modalities (sources) to investigate questions of human health that are inadequately explained by considering one image source at a time. Viewing the collection of images as multi-objects, the successful integration of multi-object data produces a sum of information greater than the individual parts, but the data complexity can hinder this integration. Each image contains structural information, indexing spatial information, or network information, and indexing connectivity among the image, reinforcing each other but challenging to merge. A Bayesian regression framework that provides inference and prediction for a multi-object outcome as a function of a scalar predictor. The framework will accommodate multiple image outcomes with different structures and jointly identify image regions associated with the scalar predictor via efficient hierarchical prior structures that scale to high-resolution image data volume. A working example is provided for the association of language comprehension scores with multi-object image data to explore the neural underpinnings of language loss in primary progressive aphasia patients.

EO244 Room 709 THE ART AND SCIENCE OF PREDICTIVE MODELING: FROM THEORY TO PRACTICE Chair: Le Zhou

E0399: **Model-based low-rank tensor clustering**
Presenter: **Qing Mai**, Florida State University, United States
Co-authors: Junge Li

Tensors have become prevalent in business applications and scientific studies. Analyzing and understanding the heterogeneity in tensor-variate observations is of great interest. A novel tensor low-rank mixture model (TLMM) IS proposed to conduct efficient estimation and clustering on tensors. The model combines the Tucker low-rank structure in mean contrasts and the separable covariance structure to achieve parsimonious and interpretable modelling. A low-rank enhanced expectation-maximization (LEEM) algorithm is developed to implement efficient computation under this model. The pseudo-E-step and the pseudo-M-step are carefully designed to incorporate variable selection and efficient parameter estimation. Numerical results in extensive experiments demonstrate the encouraging performance of the proposed method compared to popular vector and tensor methods.

E0429: **A unified framework for understanding and quantifying model privacy in adversarial machine learning**

Presenter: **Jie Ding**, University of Minnesota, United States

The security of machine learning models against adversarial attacks has become critical in modern application scenarios, such as machine-learning-as-a-service and collaborative learning. Model stealing attacks, which aim to reverse-engineer a learned model from a limited number of query-response interactions, pose a significant threat. These attacks can potentially steal proprietary models at a fraction of the original training cost. While numerous attack and defence strategies have been proposed with empirical success, most existing works are heuristic, limited in evaluation metrics, and imprecise in characterizing loss and gain. We introduce a unified conceptual framework called Model Privacy, designed to understand and quantify model stealing attacks and defence. Model Privacy captures the fundamental tradeoffs between the usability and vulnerability of a learned model's functionality. Leveraging this framework, fundamental limits are established on privacy-utility tradeoffs, and their implications are discussed. It is demonstrated that a model owner can achieve a minor utility loss by employing non-IID perturbations while obtaining a significantly larger privacy gain, a desirable property unattainable in independent data regimes. Lastly, extensive experiments are presented to corroborate the proposed framework and its effectiveness.

E0568: **Completely pivotal estimation in multivariate response linear regression models**

Presenter: **Guo Yu**, University of California Santa Barbara, United States

Despite the vast literature on sparse multivariate response linear regression models, most current methods require a known or explicit estimate of the dependence structure among the random errors. As a result, these methods hinge on computationally expensive methods (e.g., cross-validation) to determine the proper level of regularization. A completely pivotal framework for the sparse multivariate response linear regression model is proposed. Our method estimates the coefficient matrix using a model-agnostic regularization parameter that does not depend on the covariance matrix or the tail conditions of the random errors. In this sense, our proposal is completely tuning-free. Computationally, our estimator is a solution to a convex second-order cone program, which can be solved efficiently. Theoretically, the proposed estimator achieves favourable estimation error rates under mild conditions and could use a second-stage enhancement with non-convex penalties. Through comprehensive numerical studies, our method demonstrates promising statistical performance. Remarkably, our method exhibits strong robustness to violating the Gaussian assumption and significantly outperforms competing methods in heavy-tailed settings.

E1054: **Flexible regularized estimating equations: Some new perspectives**

Presenter: **Archer Yang**, McGill University, Canada

Some observations are made about the equivalences between regularized estimating equations, fixed-point problems and variational inequalities: (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 quality. Both equivalences extend to any estimating equations with convex 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, many efficient solvers exist for fixed-point problems and variational inequalities. In this regard, some efficient and scalable solvers which can deliver a hundred-fold speed improvement are applied. These connections can lead to further research in both computational and theoretical aspects of the regularized estimating equations.

Wednesday 02.08.2023 14:00 - 15:40 Parallel Session H – EcoSta2023

EV285 Room Virtual R01 APPLIED ECONOMETRICS AND STATISTICS Chair: Marica Manisera

E1032: **Finding the European crime drop using a panel data model with stochastic trends**
Presenter: **Ilka van de Werve**, Vrije Universiteit Amsterdam, Netherlands
Co-authors: Siem Jan Koopman

A new panel data model is developed with stochastically time-varying trends to empirically verify the possible existence of a European crime drop based on homicide rates from the mortality database of the World Health Organization. Different levels of pooling in the panel structures are considered, and it is shown that less pooling reveals more interesting features in the data. Extensions of the model with macroeconomic regression effects and a second trend for East-European countries are also considered. The empirical findings show strong support for a European crime drop which coincides with the well-documented US crime drop.

E1150: **A time-varying causality approach for Botswana's trade balance and its determinants**
Presenter: **Mpho Bosupeng**, Griffith University, Australia

The aim is to evaluate the causality between the trade balance and its determinants using the time-varying Granger causality approach proposed. A novel time-varying causality that is based on change detection algorithms is applied. Secondly, the new causality approach allows temporal fragilities in causal relationships to be examined through intensive subsample data analysis. The causality technique uses three algorithms to detect causality which are the forward recursive causality, rolling causality and the recursive evolving causality. Botswana depends on trading partners for the diamond trade and the time-varying causality approach will provide insights into how this reliance varies over the years. The results will assist policymakers in restructuring the economy, promoting domestic products, and measuring the effectiveness of macroeconomic policies. The time-varying Granger causality test is used to reveal the impact of domestic income, foreign income, and exchange rates on the trade balance. Botswana depends significantly on South Africa for imports and exports of diamonds to developed countries. It is expected this information will help policymakers in restructuring the economy and reaching high production capacities.

E1143: **Exploratory data analysis of innovation momentum: The application of semiconductor industry granted patents**
Presenter: **Tsung-Han Ke**, National Chi Nan University, Taiwan
Co-authors: Hung-Chun Huang, Hsin-Yu Shih

In the field of technology management, technological progress was recognized as a trajectory demonstrating technological incubation, industrial establishment, and evolution. However, the technological trajectory was constructed upon conceptual observation rather than a quantitative investigation. The aim is to utilize econometric analysis to elucidate the trajectory. The time series data of semiconductor industry granted patents are studied by using the Brock-Dechert-Scheinkman (BDS) test, augmented Dickey-Fuller (ADF) statistic, Lyapunov exponents, and Chow test. The empirical results show that the volumes of granted patents are non-linearly dependent and random, suggesting the chaotic behavior of technological innovation. Using the method of Lyapunov exponents, the nonlinear random variables are transformed into predictable stationary series. These findings could improve the statistical description of the technology trajectory.

EI004 Room 102 ADVANCES IN BAYESIAN NONPARAMETRICS Chair: Michele Guindani

E0156: **Nonparametric priors for partially exchangeable data: dependence structure and borrowing of information**
Presenter: **Igor Pruenster**, Bocconi University, Italy
Co-authors: Beatrice Franzolini, Antonio Lijoi, Giovanni Rebaudo

Partial exchangeability is the ideal probabilistic framework for analyzing data from different, though related, sources. The implications of the induced dependence structure and borrowing of information across groups are explored. These findings inspire a new general class of nonparametric priors, termed multivariate species sampling models, which is characterized by its partially exchangeable partition probability function. This class encompasses several popular dependent nonparametric priors and has the merit of highlighting their core distributional properties.

E0157: **Post-processed posteriors for high-dimensional covariances**
Presenter: **Jaeyong Lee**, Seoul National University, Korea, South

Bayesian inference of high-dimensional covariance matrices with structural assumptions, such as banded and bendable covariances, is considered, and a post-processed posterior is proposed. The post-processing of the posterior consists of two steps. In the first step, posterior samples are obtained from the conjugate inverse-Wishart posterior, which does not satisfy any structural restrictions. In the second step, the posterior samples are transformed to satisfy the structural restriction through a post-processing function. The conceptually straightforward procedure of the post-processed posterior makes its computation efficient and can render interval estimators of functionals of covariance matrices. It is shown that it has nearly optimal minimax rates for banded and bendable covariances among all possible pairs of priors and post-processing functions. Additionally, a theorem on the credible set of the post-processed posterior under the finite dimension assumption is provided. It is proved that the expected coverage probability of the (1-a)100% highest posterior density region of the post-processed posterior is asymptotically 1-a with respect to any conventional posterior distribution. It implies that the highest posterior density region of the post-processed posterior is, on average, a credible set of conventional posterior. The advantages of the post-processed posterior are demonstrated by a simulation study and a real data analysis.

E0158: **Inverse bounds and posterior contraction of the latent mixing measures**
Presenter: **Long Nguyen**, University of Michigan, United States

New results on the posterior contraction behaviour of latent mixing measures that arise in infinite and finite mixture models are presented. At the heart of this theory is a collection of inverse bounds inequalities which provide upper bounds of an optimal transport distance of mixing measures in terms of a distance of corresponding data population distributions. For infinite mixtures, existing posterior contraction results measured in terms of the Wasserstein metrics are quite rare and confined to location mixtures. Nonetheless, more can still be said for Gaussian mixtures in particular, where it can be shown that while the overall convergence behaviour is slow, the parameters in the outlier regions of the parameter space converge almost at polynomial rates. This is possible by employing a generalized notion of optimal transport distance known as the Orlicz-Wasserstein metric. This is a welcome result for the practitioners of (infinite) Gaussian mixtures who wish to be able to interpret the model parameters efficiently. For finite mixtures where the number of components is unknown, a rather general theory has emerged based on a novel notion of strong identifiability with respect to any class of test functions subject to suitable conditions. This theory allows us to analyze a broader class of mixtures than has been considered before.

EO171 Room 02 JOINT MODELLING OF MULTI-OUTCOME DATA Chair: Christiana Charalambous

E0280: **A general joint latent class model of longitudinal and survival data with covariance modelling**
Presenter: **Ruoyu Miao**, University of Cambridge, United Kingdom
Co-authors: Christiana Charalambous

Based on the proposed time-varying joint latent class model (JLCM), the heterogeneous random covariance matrix can also be considered, which regression submodel for the variance-covariance matrix of the multivariate latent random effects can be added to the joint latent class model. A general JLCM with heterogeneous random-effects modelling is a natural extension of the time-varying JLCM, which consists of the linear and the log link functions to model the covariance matrices as the variance-covariance regression submodel based on the modified Cholesky decomposition, longitudinal submodel, survival submodel as well as the membership probability. The covariance modelling enables us to determine any effects of covariates on the association between the longitudinal and survival processes while also allowing each subject's group classification to change over time. The assumption can also be tested by adding the regression model and the homogeneous random effects. The Bayesian approach will be used to do the estimation. DIC value is the criteria to decide the optimal k value. Our general JLCM is illustrated on a real data set of aids study in which the prospective accuracy of our proposed JLCM is of interest, as did the dynamic predictions for time-to-death in the joint model using the longitudinal CD4 cell count measurements.

E0717: **A two-level copula joint model for joint analysis of longitudinal and competing risks data**
Presenter: **Thierry Chekouo**, University of Minnesota, United States

A two-level copula joint model is proposed to analyze clinical data with multiple disparate continuous longitudinal outcomes and multiple event times in the presence of competing risks. At the first level, a copula to model the dependence between competing latent event times, in the process of constructing the submodel for the observed event time is used, and the Gaussian copula is employed to construct the submodel for the longitudinal outcomes that account for their conditional dependence; these submodels are glued together at the second level via the Gaussian copula to construct a joint model that incorporates conditional dependence between the observed event-time and the longitudinal outcomes. To have the flexibility to accommodate skewed data and examine possibly different covariate effects on quantiles of a non-Gaussian outcome, linear quantile mixed models are proposed for the continuous longitudinal data. A Bayesian framework is adopted for model estimation and inference via Markov Chain Monte Carlo sampling. The performance of the Copula joint model is examined through a simulation study, and it is shown that the proposed method outperforms the conventional approach assuming conditional independence with smaller biases and better coverage probabilities of the Bayesian credible intervals. Finally, an analysis of clinical data is carried out on renal transplantation for illustration.

E0983: **Modeling past event feedback through biomarker dynamics in the multistate event analysis for cardiovascular disease data**
Presenter: **Chuoxin Ma**, Beijing Normal University-Hong Kong Baptist University United International College, China
Co-authors: Jianxin Pan

In cardiovascular studies, ordered multiple events along disease progression are observed, which are essentially a series of recurrent events and terminal events with competing risk structures. One of the main interests is to explore the event-specific association with the dynamics of longitudinal biomarkers. A new statistical challenge arises when the biomarkers carry information from the past event history, providing feedback for the occurrences of future events and particularly when these biomarkers are only intermittently observed with measurement errors. A novel modelling framework isis proposed where the recurrent events and terminal events are modelled as multistate processes, and random effects models describe the longitudinal covariates that account for event feedback. Flexible models with semiparametric coefficients are adopted, considering the nature of long-term observation in cardiac studies. A one-step estimator of the regression coefficients is developed to improve computation efficiency, and their asymptotic variances for the computation of the confidence intervals are derived based on the proposed asymptotically unbiased estimating equation.

E1130: **Joint models for multi-outcome data and covariance structures via a Bayesian approach**
Presenter: **Christiana Charalambous**, University of Manchester, United Kingdom
Co-authors: Ruoyu Miao

In risk prediction for cardiovascular disease (CVD), where a risk factor such as systolic blood pressure (SBP) is volatile, giving high within-subject variability, then correctly modelling that variability could offer further improvements compared to the classic joint model for SBP and CVD. Motivated by this example, joint models are proposed for the survival outcome (time to CVD) as well as both the mean and variance of the longitudinal outcome (SBP). These models are linked via heterogeneous random effects sharing the same distribution, allowing us to capture the pairwise associations between the three outcomes through the random effects covariance matrix. Both the modified Cholesky and Hypersphere decompositions are considered to reparameterise the conditional covariance of the longitudinal response and employ a Bayesian approach for estimation. The performance of the proposed approach is demonstrated via simulation and application to the Systolic Blood Pressure Intervention Trial (SPRINT) dataset.

EO251 Room 03 ADVANCED STATISTICAL METHODS FOR BIOMEDICAL DATA Chair: Eunjee Lee

E0649: **A fast and powerful spatial-extent inference for testing variance components in reliability and heritability studies**
Presenter: **Jun Young Park**, University of Toronto, Canada

Clusterwise inference is a popular approach in neuroimaging to increase sensitivity, but most existing methods are currently restricted to the General Linear Model (GLM) for testing mean parameters. Statistical methods for testing variance components, which are critical in neuroimaging studies that involve estimation of narrow-sense heritability or test-retest reliability, are seriously underdeveloped due to methodological and computational challenges, which would potentially lead to low power. To fill this gap, a fast and powerful test for variance components called CLEAN-V ("CLEAN" for testing "V"ariance components) is proposed. CLEAN-V models imaging data's global spatial dependence structure and compute a locally powerful variance component test statistic by data-adaptively pooling neighbourhood information. Permutations achieve correction for multiple comparisons to accurately control family-wise error rate (FWER). Through analysis of task-fMRI data from the Human Connectome Project (HCP) across five tasks and comprehensive data-driven simulations, it is shown that CLEAN-V outperforms existing methods in detecting test-retest reliability and narrow-sense heritability with significantly improved power, with the detected areas aligning with activation maps. The computational efficiency of CLEAN-V also speaks of its practical utility, and it is available as an R package.

E0703: **A deep attention LSTM embedded aggregation network for multiple histopathological images**
Presenter: **Sunghun Kim**, Chungnam National University, Korea, South
Co-authors: Eunjee Lee

Recent computer vision and neural network advancements have facilitated medical imaging survival analysis for various medical applications. However, challenges arise when patients have multiple images from multiple lesions, as current deep-learning methods can provide multiple survival predictions for each patient, complicating result interpretation. To address this issue, a deep-learning survival model that can provide accurate predictions at the patient level was developed. A Deep Attention Long Short-Term Memory Embedded Aggregation Network (DALAN) is proposed for histopathology images, designed to perform feature extraction and aggregation of lesion images simultaneously. This design enables the model to learn imaging features from lesions efficiently and aggregate lesion-level information to the patient level. DALAN comprises a weight-shared CNN, attention layers, and LSTM layers. The attention layer calculates the significance of each lesion image, while the LSTM layer combines the weighted information to produce an all-encompassing representation of the patient's lesion data. DALAN was evaluated against several naive aggregation methods on simulated and real datasets in terms of the c-index. The results showed that DALAN outperformed the competing methods on the MNIST and Cancer dataset simulations. On the real TCGA dataset, DALAN also achieved a higher c-index of 0.803 compared to the naive methods and the competing models.

E0779: **A fully Bayesian tensor basis model for multi-subject task fMRI data**
Presenter: **Michelle Miranda**, University of Victoria, Canada
Co-authors: Jeffrey Morris

Task-evoked functional magnetic resonance imaging (fMRI) studies are a powerful tool for understanding human sensory, cognitive, and emotional processes. A Bayesian approach is introduced to analyze task fMRI data that simultaneously detects activation signatures and background connectivity. The joint modelling involves a subjective-specific tensor spatial-temporal basis strategy that enables scalable computing yet captures spatial correlation from nearby voxels, distant ROIs, and long-memory temporal correlation. The spatial basis involves a composite hybrid transform with two levels: the first accounts for within-ROI correlation, and the second is a between-ROI distant correlation. The proposed basis space regression modelling strategy increases sensitivity for identifying activation signatures, partly driven by the induced background connectivity that can be summarized to reveal biological insights. This strategy leads to computationally scalable fully Bayesian inference at the voxel or ROI level that adjusts for multiple testing. Moreover, a joint, fully Bayesian multi-subject model is introduced and used to gain insights into the working memory task of the Human Connectome Project.

E0796: **On optimal biomarker cutoffs accounting for misclassification costs in diagnostic trilemmas**
Presenter: **Leonidas Bantis**, University of Kansas Medical Center, United States
Co-authors: John Tsimikas

The ROC surface is an appropriate tool for assessing the overall accuracy of a marker employed under trichotomous settings. A decision/classification rule is often based on the so-called Youden index and its three-dimensional generalization. However, both the clinical and the statistical literature have not paid the necessary attention to the underlying false classification (FC) rates that are of equal or even greater importance. We provide a framework to make inferences around all classification rates as well as relevant comparisons. We also accommodate the underlying misclassification costs associated with the false classification rates. We explore the trinormal model, flexible models based on power transformations, and robust non-parametric alternatives. We evaluate our approaches through extensive simulations and illustrate them using data from a study that involves patients with pancreatic cancer.

EO307 Room 04 SURVIVAL ANALYSIS AND BIOMEDICAL STATISTICS Chair: Dongdong Li

E0245: **The mean residual life model for the survival data with covariate measurement errors: Application to stage 5 CKD data**
Presenter: **Chyong-Mei Chen**, Institute of Public Health, Taiwan

The mean residual life regression model with covariate measurement errors is considered. In the whole cohort, the surrogate variable of the error-prone covariate is available for each subject. In contrast, the instrumental variable (IV), which is related to the underlying true covariates, is measured only for some subjects, the calibration sample. Without specifying distributions of measurement errors and assuming that the IV is missing at random, the calibration and cohort estimators of the regression parameters are proposed by solving estimation equations (EEs) developed from the calibration and cohort samples, respectively. A synthetic estimator is derived by synthesizing the EEs based on the generalized method of moments to improve estimation efficiency. The large sample properties of the three proposed estimators are derived, and their finite sample performance is evaluated via simulation studies. Simulation results show that the cohort and synthetic estimators outperform the IV calibration estimator, and the relative efficiency of the cohort and synthetic estimators depends on the missing rate of IV. The proposed method is illustrated by application to data from patients with stage 5 chronic kidney disease in Taiwan.

E0673: **Mean residual life based illness-death model for semicompeting risks data**
Presenter: **Liming Xiang**, Nanyang Technological University, Singapore

Semicompeting risks data are available in many biomedical studies, where some nonterminal event (e.g., disease progression) is of interest and subject to censoring by the terminal event (e.g., death). The illness-death model is commonly used for the analysis of such data. It is proposed to formulate the effects of covariates on the endpoints through semiparametric mean residual life regression models with shared frailty under the illness-death model framework. Novel estimating equations are developed based on a penalized quasi-likelihood incorporating the inverse probability of censoring weights to adjust for possible dependent censoring. Unlike the usual illness-death model assuming a gamma frailty, the proposed inference procedure requires no distributional assumption for frailty. Under some regularity conditions, it is shown that the resulting parameter estimators are consistent and asymptotic normal. Simulation results demonstrate that the method performs well in various realistic settings. The usefulness of the method is further illustrated via the analysis of a real data example.

E0642: **Bayesian ridge regression for survival data based on a vine copula based prior**
Presenter: **Takeshi Emura**, The Institute of Statistical Mathematics, Japan
Co-authors: Hirofumi Michimae

Ridg regression with the Cox model is regarded as a Bayesian estimator with a multivariate normal prior. The vine copula-based priors are proposed for Bayesian Cox ridge estimators under the proportional hazards model. The vine copula allows the tail dependence that is not possible by multivariate normal priors. The semiparametric Cox models are built on the posterior density under two likelihoods: Cox's partial likelihood and the full likelihood under the gamma process prior. It is also shown via simulations and a data example that the Archimedean vine copula priors (the Clayton and Gumbel copula) are superior to the multivariate normal prior and the Gaussian copula prior.

E0932: **Estimating optimal treatment regimes in semi-supervised framework**
Presenter: **Mengjiao Peng**, East China Normal University, China, China

Finding the optimal individualized treatment rule has been studied intensively in the literature, with important applications in practice. The problem of estimating the optimal treatment regime in a semi-supervised learning setting is considered, where a very small proportion of the entire set of observations is labelled with the true outcome but features predictive of the outcome are available among all observations. A model-free robust inference approach for optimal treatment regimes is proposed with the aid of the unlabeled data with only covariate information to improve estimation efficiency. The proposed estimation of OPT primarily involves a flexible nonparametric imputation by single index kernel smoothing which works well even for high-dimensional covariates, and a follow-up estimation for optimal treatment regime based on concordance-assisted learning, including optimization of the estimated concordance function up to a threshold and finding the optimal threshold to maximize the inverse propensity score weighted (IPSW) estimator of the value function. Moreover, when the propensity score function is unknown, a doubly robust estimation method is developed under a class of monotonic index models. The estimators are shown to be consistent and asymptotically normal. Simulations exhibit the efficiency and robustness of the proposed method compared to existing approaches in finite samples.

EO091 Room Virtual R02 STATISTICAL METHODS AND DEPENDENCE IN SPACE AND/OR TIME Chair: Moritz Jirak

E1084: **Dimension-free rates of bootstrap approximation for spectral statistics in high-dimensional PCA**
Presenter: **Miles Lopes**, UC Davis, United States

In the context of principal components analysis (PCA), the bootstrap is commonly applied to solve a variety of inference problems, such as constructing confidence intervals for the eigenvalues of the population covariance matrix. However, when the data are high-dimensional, there are relatively few theoretical guarantees that quantify the performance of the bootstrap. A number of recent results will be discussed that establish rates of bootstrap approximation for statistics arising in high-dimensional PCA. Examples include the leading eigenvalues of the sample covariance matrix, as well as the operator norm error of this matrix with respect to the population covariance matrix. Notably, in settings where the population eigenvalues exhibit a decaying structure, the rates of bootstrap approximation are dimension-free.

E1162: **A kernel-based analysis of Laplacian eigenmaps**
Presenter: **Martin Wahl**, Bielefeld University, Germany
Co-authors: Martin Wahl

Laplacian eigenmaps and diffusion maps are nonlinear dimensionality reduction methods that use the eigenvalues and eigenvectors of normalized graph Laplacians. From a mathematical perspective, the main problem is understanding these empirical Laplacians as spectral approximations of the underlying Laplace-Beltrami operator. Laplacian eigenmaps are studied through the lens of kernel principal component analysis (PCA). This leads to novel points of view and allows us to leverage methods developed for PCA in infinite dimensions.

E1265: **Sharp adaptive similarity testing with pathwise stability for ergodic diffusions**
Presenter: **Johannes Brutsche**, University of Freiburg, Germany
Co-authors: Angelika Rohde

Within the nonparametric diffusion model, multiple tests are developed to infer the similarity of an unknown drift b to some reference drift b_0: At prescribed significance, those regions are simultaneously identified where violation from similarity occurs, without a priori knowledge of their number, size and location. This test is shown to be minimax-optimal and adaptive. At the same time, the procedure is robust under small deviations from Brownian motion as the driving noise process. A detailed investigation for fractional driving noise, which is neither a semimartingale nor a Markov process, is provided for Hurst indices close to the Brownian motion case.

E1278: **Weak dependence and optimal quantitative self-normalized central limit theorems**
Presenter: **Moritz Jirak**, University of Vienna, Austria

Consider a stationary, weakly dependent sequence of random variables. Given that a CLT holds, how should the long-run variance be estimated? This problem has been studied for decades, and prominent proposed solutions have been given. Using the proximity of the corresponding normal distribution as a quality measure, optimal solutions and why previous proposals are not optimal in this context discussed. The setup contains many prominent dynamical systems and time series models, including random walks on the general linear group, products of positive random matrices, functionals of Garch models of any order, functionals of dynamical systems arising from SDEs, iterated random functions and many more.

EO150 Room 201 CAUSAL INFERENCE IN OBSERVATIONAL STUDIES Chair: Yeonseung Chung

E1029: **Bayesian additive regression trees model for high-dimensional potential confounders**
Presenter: **Chanmin Kim**, SungKyunKwan University, Korea, South

A solution to the increasing challenge faced in the analysis of observational studies is offered, which involves identifying the covariates required to establish the assumption of ignorable treatment assignment for causal effect estimation. The proposal adopts a Bayesian nonparametric approach that addresses this challenge in three ways. First, it prioritizes the inclusion of adjustment variables based on established confounder selection principles. Second, it enables estimating causal effects by accounting for complex relationships among confounders, exposures, and outcomes. Finally, it provides causal estimates that consider the uncertainty in the confounding nature. The method involves using multiple Bayesian Additive Regression Tree models that share a prior distribution, accumulating posterior selection probability to covariates associated with both the exposure and the outcome of interest. Several simulation studies demonstrate that the proposed method performs well relative to other similar methods in various scenarios. The approach is applied to examine the causal effect of SO2 emissions from coal-fired power plants on ambient air pollution concentrations, providing strong evidence of the causal relationship between SO2 emissions and ambient particulate pollution over adjacent years.

E1117: **Differential recall bias in self-reported risk factors in observational studies**
Presenter: **Suhwan Bong**, Seoul National University, Korea, South
Co-authors: Kwonsang Lee, Francesca Dominici

Observational studies are typically used to estimate the effect of exposures on outcomes. Treatment effect estimation can only be unbiased if the exposure is correctly measured. Recall bias is one of the common reasons for exposure misclassification. Recall bias can occur when study subjects do not remember previous events accurately or omit details. First, the estimand of interest is identified: the average treatment effect (ATE) in the presence of recall bias. Several estimation approaches for the ATE are also developed. These methods are then implemented in simulations to demonstrate their performance in different model misspecification scenarios. Finally, the proposed framework is applied to an observational study, estimating the effect of childhood physical abuse on adulthood mental health.

E1198: **Causal clustering**
Presenter: **Kwangho Kim**, Korea University, Korea, South
Co-authors: Edward Kennedy, Larry Wasserman

Causal effects are often characterized by population effects, which can give an incomplete picture when the treatment effect within subpopulations varies considerably from the population effect. As the subgroup structure is usually unknown, identifying and estimating subpopulation effects is relatively more challenging than population-level effects. Causal Clustering, a new set of methods for exploring the heterogeneity of treatment effects, leveraging tools from clustering analysis, is developed. First, an efficient way is developed to uncover subgroup structure by harnessing widely-used clustering methods. Specifically, it is shown that k-means, density-based, and hierarchical clustering algorithms can be successfully adopted into our framework via plug-in type estimators. Next, for the k-means causal clustering, a specially bias-corrected estimator based on nonparametric efficiency theory is developed, which attains fast convergence rates and asymptotic normality to the true cluster centres under weak nonparametric conditions. This requires novel techniques due to the particular form of the non-smooth k-means risk. Novel tools especially useful for modern outcome-wide studies with many treatment levels are derived. Importantly, it is also discussed how these methods can be extended to clustering with generic pseudo-outcomes: e.g., partially observed outcomes or unknown functionals. Finally, the methods are illustrated via simulation studies and real data analyses.

E1308: **Causal inference in environmental epidemiology research with large-size retrospective cohort data**
Presenter: **Whanhee LEE**, Pusan National University, Korea, South

The causal inference has been raised as a very important topic in environmental epidemiology research, especially in relation to air pollution. Including the U.S. Environmental Protection Agency (EPA); although many regulatory authorities have tried to establish their air pollution standards based on causal evidence, the related research is limited, still. To address this limitation, recently, several studies have attempted to provide generalizable air pollution risk estimates based on large retrospective cohort data with causal inference analytic methods. In particular, the national health insurance system-based claim cohort has been widely used in this research field. Thus, the current stage and emerging methodological/analytic issues in epidemiological studies are addressed to find the causal association between air pollution and health outcome with the health insurance cohort data.

EO145 Room 203 RECENT ADVANCES ON QUANTILE AND TAIL ANALYSIS Chair: Qian Xiao

E0230: **Whittle estimation based on the extremal spectral density of a heavy-tailed random field**
Presenter: **Yuwei Zhao**, Fudan University, China

A strictly stationary random field is considered on the two-dimensional integer lattice with varying marginal and finite-dimensional distributions. Exploiting the regular variation, the spatial extremogram, which is defined, takes into account only the largest values in the random field. This extremogram is a spatial autocovariance function. The corresponding extremal spectral density and its estimator, the extremal periodogram, are described. Based on the extremal periodogram, the Whittle estimator for suitable classes of parametric random fields is considered, including the Brown-Resnick random field and regularly varying max-moving averages.

E0803: **Comparing time varying regression quantiles under shift invariance**
Presenter: **Weichi Wu**, Tsinghua University, China
Co-authors: Subhra Sankar Dhar

The aim is to investigate whether time-varying quantile regression curves are the same up to the horizontal shift or not. The errors and the covariates involved in the regression model are allowed to be locally stationary. This issue is formalized in a corresponding non-parametric hypothesis testing problem, and an integrated-squared-norm based test (SIT) and a simultaneous confidence band (SCB) approach are developed. The asymptotic properties of SIT and SCB under null and local alternatives are derived. Moreover, the asymptotic properties of these tests are also studied when the compared data sets are dependent. Then valid wild bootstrap algorithms are developed to implement SIT and SCB. Furthermore, the usefulness of the proposed methodology is illustrated by analyzing simulated and real data related to the COVID-19 outbreak.

E0993: **Panel quantile regression for extreme risk**
Presenter: **Yanxi Hou**, Fudan University, China

Panel quantile regression models play an essential role in real finance, econometrics, insurance, and risk management applications. However, direct estimates of the extreme conditional quantiles may lead to unstable results due to data sparsity on the far tail. Moreover, the presence of individual effects in panel quantile regressions complicates the inference for high quantiles. A two-stage method is proposed to estimate/predict the high conditional quantiles. The intermediate quantiles are first predicted according to panel quantile regressions, and the extreme quantiles are obtained by extrapolating the intermediate ones in the second stage. The asymptotic properties of the prediction method rely on a set of second-order conditions for heteroscedastic extremes. A metric called Average Absolute Relative Error is used to evaluate the prediction performance of high conditional quantiles over different cross-sections. The asymptotic distributions of the metric for both intermediate and extreme quantiles are studied. The two-stage prediction's finite sample performance is demonstrated, compared to the direct prediction for extreme conditional quantiles. Finally, the two-stage method is applied to the macroeconomic and housing price data, and strong evidence of housing bubbles and common economic factors is found.

E1003: **Systemic and systematic risks driven marginal expected shortfall**
Presenter: **Deyuan Li**, Fudan University, China

Marginal expected shortfall (MES) may be referred to the expected loss of a particular equity during the occurrence of a system-wide stress event (SWSE) or equivalently systemic risk and/or systematic risk (SRSR) in a system. The existing MES literature only considers the occurrence of SWSE being an index return dropping below a prescribed threshold, which can be insufficient in characterizing an SWSE, and hence can be largely underestimating the expected loss of the equity. An SWSE is innovatively defined as a representative index return dropped below a prescribed threshold or the worst performed individual equity's return dropped below a prescribed threshold, extending the MES to the innovative systemic and systematic risks driven marginal expected shortfall (SYS2MES). Estimators for SYS2MES are constructed, and their asymptotic theories are established within the multivariate extreme value theory framework. The results cover both the tail-dependent and tail-independent cases and thus can be applied to a wide range of models. The finite sample performance of the estimators is investigated in a simulation study. Applying SYS2MES to Dow Jones' 30 stocks led to better and more meaningful results than the original MES. The new results make invaluable market risk measurements and management.

EO223 Room 503 BAYESIAN MODELLING WITH MIXTURE MODELS Chair: Cheng Li

E0618: **Flexible modelling of heterogeneous populations of networks: A Bayesian nonparametric approach**
Presenter: **Francesco Barile**, Bicocca University, Italy
Co-authors: Bernardo Nipoti, Simon Lunagomez

The increasing availability of multiple network data has been calling for the development of statistical models for heterogeneous populations of networks. A popular approach to the problem of clustering multiple network data uses distance metrics that measure the similarity among networks based on some of their global or local characteristics. In this context, a novel Bayesian nonparametric approach is proposed to model undirected labelled graphs sharing the same set of vertices, which allows us to identify clusters of networks characterized by similar patterns in the connectivity of nodes. Our construction relies on the definition of a location-scale Dirichlet process mixture of centred Erdos-Renyi (CER) kernels. A unique mode or network representative and a univariate measure of dispersion around the mode conveniently parametrize the CER kernel function. An efficient Markov chain Monte Carlo scheme is proposed to carry out posterior inference and conveniently cluster the multiple network data. The number of clusters in the population is not set a priori but inferred from the data. The performance of our approach is investigated by means of an extensive simulation study and illustrated with the analysis of a dataset on brain networks.

E0860: **Hierarchically dependent mixture hazard rates for modelling competing risks**
Presenter: **Claudio Del Sole**, Bocconi University, Italy
Co-authors: Antonio Lijoi, Igor Pruenster

A popular approach in Bayesian modelling of partially exchangeable data consists in imposing hierarchical nonparametric priors, which induce dependence across groups of observations. In survival analysis, hierarchies of completely random measures have been successfully exploited as mixing measures to model multivariate dependent mixture hazard rates, leading to a posterior characterization that may accommodate censored observations. Such a framework can be easily adapted to a competing risks scenario, in which groups correspond to different diseases affecting each individual. In this case, the multivariate construction acts at a latent level, as only the minimum time-to-event and the corresponding cause of death are observed. The posterior hierarchy of random measures and the posterior survival and cause-specific incidence function estimates are explicitly described conditionally on a suitable latent partition structure that fits the Chinese restaurant franchise metaphor. Marginal and conditional sampling algorithms are also devised and tested on synthetic datasets. The performances of this proposal are finally compared with those of its non-hierarchical counterpart, which models the hazard rate of each disease independently: leveraging the information borrowed from other groups, the hierarchical construction is empirically shown to recover the shape of the incidence functions more efficiently, in the presence of proportional hazards.

E0882: **Bayesian nonparametric modelling of latent partitions via Stirling-gamma priors**
Presenter: **Alessandro Zito**, Duke University, United States
Co-authors: Tommaso Rigon, David Dunson

The Dirichlet process (DP) has received much attention in recent decades as an effective tool for clustering and density estimation. However, DP mixtures are particularly sensitive to the value of the precision parameter, which must be chosen carefully to prevent over-clustering. Moreover, common choices of priors for the precision, such as the gamma distribution, induce an analytically intractable prior over the associated number of clusters. A class of priors is introduced for the precision parameter that instead makes the induced prior over the number of clusters tractable and approximately distributed as a Negative Binomial. The prior belongs to the novel class of Stirling-gamma distributions, which are flexible and easily sampled from. It has been shown how certain choices of the hyperparameters of the Stirling-gamma allow obtaining conjugacy to the law of the random partition generated by a DP and the number of associated clusters therein. This leads to a very interpretable prior specification, simplifying both prior elicitation and posterior computations. The resulting marginal process is framed within the larger class of Gibbs-type partition models. The conjugate case also shows how the Stirling-gamma allows for the borrowing of information across multiple observed partitions.

E0886: **Joint species distribution modeling with mixture models**
Presenter: **Ching-Lung Hsu**, Duke University, United States
Co-authors: Tommaso Rigon, David Dunson

Community ecology seeks to understand the interactions between species and their driving environmental factors. Joint species distribution models are increasingly studied and applied to ecological data for estimating species associations and the predictive power in future samples. A Bayesian mixture model is proposed for co-occurrence probabilities allowing the discovery of unseen species. The results are leveraged from previous work, and the predictive formulas are derived for species discovery in this setting. It is shown that asymptotically the model is a mixture of a two-parameter Indian buffet process. A simple Gibbs sampler is developed for posterior computation. As an implementation, the model is applied to the Guelph Arthropod dataset.

EO046 Room 506 STATISTICAL LEARNING IN FINANCE Chair: Li-Hsien Sun

E0734: **Kullback-Leibler divergence and Akaike information criterion in general hidden Markov models**
Presenter: **Chu-Lan Kao**, National Yang Ming Chiao Tung University, Taiwan
Co-authors: Tianxiao Pang, Cheng-Der Fuh

To characterize the Kullback-Leibler divergence and Fisher information in general parametrized hidden Markov models, first, it is shown that the log-likelihood and its derivatives can be represented as an additive functional of a Markovian iterated function system, and then provide explicit characterizations of these two quantities through this representation. Moreover, it is shown that Kullback-Leibler divergence can be locally approximated by a quadratic function determined by the Fisher information. Results relating to the Cramer-Rao lower bound and the Hajek-Le Cam local asymptotic minimax theorem are also given. As an application of the results, a theoretical justification for using the Akaike information criterion (AIC) model selection in general hidden Markov models is provided. Last, three concrete models are studied: a Gaussian vector autoregressive-moving average model of order (p,q), recurrent neural networks, and a temporally restricted Boltzmann machine to illustrate the theory.

E1042: **Change point detection through copula-based Markov models**
Presenter: **Li-Hsien Sun**, National Central University, Taiwan
Co-authors: Ming-Hua Hsieh, Dong-Hua Kuo

Time series analysis is critical in various fields such as finance, industry, and biology. However, due to the possibility of the structure change, problems, such as loss or damage, can be expected (e.g., the stock market during the financial crisis in 2008 and COVID-19 in 2020). Hence, the corresponding change point for structural change is worth studying. In order to detect the changepoint online for time series data or correlated data, the model is proposed for online changepoint detection via copula-based Markov models where the time serial data is described by copula-based Markov model and the change-point detection based on the run length distribution using the Bayesian approach. Finally, the performance of the proposed method is illustrated through numerical and empirical studies.

E1095: **Fuzzy-based evaluation system for position overweighting mechanism**
Presenter: **ChiFang Chao**, National Taipei University of Technology, Taiwan
Co-authors: Mu-En Wu, Ming-Hua Hsieh

A novel fuzzy system is proposed for evaluating overweighting position mechanisms and providing a precise quantification of their suitability. Historical data from the Taiwan Futures Exchange (TAIFEX) Futures was used to generate corresponding return probability distributions for evaluating the performance of four trading logics with position overweighting mechanisms. The parameters with the highest QIs were used to establish the optimal portfolio, with the weight determined by the degree of fuzzy using the softmax fuzzy set to transform the QI. The optimal portfolio was then compared to 10,000 corresponding random portfolios generated by Monte Carlo simulation. Results indicate that the proposed system accurately quantified suitability. The optimal portfolio generated for high-frequency data exhibited excellent performance and outperformed over 80% random portfolios. The system was also applied to S&P 500 futures, crude oil futures, and soybean futures, with results indicating that the system can distinguish 75% suitability and, on average, outperforms 94% of randomly generated portfolios.

E1059: **Estimation of threshold boundary regression models**
Presenter: **ChihHao Chang**, National University of Kaohsiung, Taiwan
Co-authors: Takeshi Emura, Shih-Feng Huang

The threshold boundary regression (TBR) model is considered for sample splitting. The TBR model accommodates covariates in both the regression and threshold functions. The threshold function is allowed to be a nonlinear function of multiple covariates, constituting a hyperplane to describe data dynamics in two different states. TBR-WSVM, a two-stage method, is proposed that incorporates the weighted support vector machine (WSVM) and least-squares (LS) methods to estimate the TBR model. Under regularity conditions, the consistency of the TBR-WSVM estimators with their optimal convergence rates is evaluated. Several simulation experiments are conducted to investigate the finite sample performance of the TBR-WSVM estimator. Compared with two recently proposed methods, TBR-WSVM enjoys three advantages: (i) threshold parameters need not be prefixed with nonzero values, (ii) threshold parameter ranges need not be specified, and (iii) the threshold boundary can be non-linearly estimated. Finally, the TBR model is applied to real data analysis.

EO055 Room 604 ECONOMETRICS AND STATISTICS FOR THE DIGITAL ASSET ECONOMY Chair: Jeffrey Chu

E0745: **A time series approach to explainability for neural netswith applications to risk-management and fraud detection**
Presenter: **Branka Hadji Misheva**, BFH, Switzerland

Artificial intelligence (AI) is creating one of the biggest revolutions across technology-driven application fields. The finance sector offers many opportunities for significant market innovation, yet the broad adoption of AI systems heavily relies on our trust in their outputs. Trust in technology is enabled by understanding the rationale behind the predictions made. In other words, it needs to ensure that values and domain knowledge are reflected in the algorithms' outcomes. To this end, the concept of eXplainable AI (XAI) emerged, introducing a suite of techniques attempting to explain to users how complex models arrived at a certain decision. Even though many of the classical XAI approaches can lead to valuable insights about the models' inner workings, in most cases, these techniques are not tailored for time series applications due to the presence of possibly complex and non-stationary dependence structure of the data. A generic XAI technique for deep learning methods (DL) is proposed, which preserves and exploits the natural time ordering of the data by introducing a family of so-called explainability (X-)functions. This concept bypasses severe identifiability issues, related among others, to profane numerical optimization problems, and it promotes transparency by means of intuitively appealing input-output relations ordered by time.

E0763: **Zigzag filtration curve based supra-hodge convolution networks for dynamic ethereum token networks forecasting**
Presenter: **Yuzhou Chen**, Temple University, United States

Graph neural networks (GNNs) offer a new powerful alternative for multivariate time series forecasting, demonstrating remarkable success in various spatiotemporal applications, from urban flow monitoring systems to health care informatics to financial analytics. Yet, such GNN models pre-dominantly capture only lower order interactions, that is, pairwise relations among nodes, and also largely ignore intrinsic time-conditioned information on the underlying topology of multivariate time series. To address these limitations, a new time-aware GNN architecture which amplifies the power of the recently emerged simplicial neural networks with a time-conditioned topological knowledge representation in the form of zigzag persistence, is proposed. That is, the new approach, Zigzag Filtration Curve based Supra-Hodge Convolution Networks (ZFC-SHCN), is built upon the two main components – a new highly computationally efficient zigzag persistence curve and a new temporal multiplex graph representation module for learning higher-order network interactions. Theoretical properties of the proposed time-conditioned topological knowledge representation and extensive validate the new time-aware ZFC-SHCN model in conjunction with time series forecasting on Ethereum blockchain datasets are discussed. The experiments demonstrate that the ZFC-SHCN achieves state-of-the-art performance with lower requirements on computational costs.

E0791: **The financial impact of war on cryptocurrencies**
Presenter: **Jeffrey Chu**, Renmin University of China, China

Over the past decade, new and old cryptocurrencies have been exposed to a wide variety of significant global events, such as financial crises, rising inflation, booms and recessions, and, most recently, the coronavirus (COVID-19) pandemic. Up until 2022, cryptocurrencies had never witnessed a military conflict and simultaneously played a significant role. This all changed in February 2022 with the Russia-Ukraine conflict. The key question is: how has the conflict impacted cryptocurrency markets? The impact of war-related events on the cryptocurrency markets through an event-study approach and how the cryptocurrency markets have evolved throughout this period through a network graph approach are discussed.

E0838: **Stylized facts of decentralized finance (DeFi)**
Presenter: **Stephen Chan**, American University of Sharjah, United Arab Emirates

Decentralized Finance (DeFi) represents an emerging sector within the cryptocurrency space. DeFi is currently one of the most groundbreaking and disruptive technologies impacting centralized finance systems, bringing with it many distinctive features and huge potential. Stylized facts on DeFi are presented, and light is shed on the broader empirical features of market efficiency, volatility clustering, leverage effects, and the return volume relationship of this market.

EO134 Room 606 RECENT DEVELOPMENTS IN TIME-SERIES AND ECONOMETRICS Chair: Shih-Feng Huang

E0316: **An asymptotic behaviour of a finite-section of the optimal causal filter**
Presenter: **Junho Yang**, Academia Sinica, Taiwan

An L_1-bound between the coefficients of the optimal causal filter applied to the data-generating process and its approximation based on finite sample observations is derived. The data-generating process is assumed to be second-order stationary with either short or long-memory autocovariance. First, to obtain the L_1-bound, an exact expression of the causal filter coefficients and their approximation in terms of the absolute convergent series of the multistep ahead infinite and finite predictor coefficients, respectively, are provided. Then, a so-called uniform-type Baxter's inequality is proved to obtain a bound for the difference between the two multistep ahead predictor coefficients (under short and memory time series). The L_1-approximation error bound of the causal filter coefficients can be used to evaluate the quality of the time series predictions through the mean squared error criterion.

E0488: **LIMOS–LightGBM interval Merton one-period-portfolio selection**
Presenter: **Liang-Ching Lin**, National Cheng Kung University, Taiwan
Co-authors: Sz-Wei Charng

The modern portfolio theory can assist us in allocating wealth to risky and risk-free assets reasonably by using some statistical methods. The aim is to focus on evolving Merton's portfolio problem. Instead of the conventional parameter estimations based on only the closing prices, the opening, high, low, and closing prices are included to enlarge the database as much as possible to make the parameter estimations much more accurate. Furthermore, a weighted arithmetic mean of estimations obtained is considered from different lengths of training datasets to stabilize the estimators in which the weights are evaluated by using the least-squares method. In addition, the LightGBM is used to predict the transaction directions, and not only the prices as tradition and also many statistics are included to be the features. In real data analysis, it is demonstrated the usefulness of combining the aforementioned methods by showing the portfolio profits of selecting ten stocks in 2018 and 2019. The results particularly show the superiority of the proposed strategy over the conventional method: the profits are almost positive and have around 32% to 72% annually.

E0521: **Consistent autoregressive spectral estimates under GARCH-type noises**
Presenter: **Hsin-Chieh Wong**, National Taipei University, Taiwan
Co-authors: Ching-Kang Ing, Wen-Jen Tsay

The semiparametric estimation of the spectral density function of a stationary time series driven by a general class of noise with conditional heteroskedasticity is considered. First, it is shown that the ordinary least squares (OLS) based autoregressive regression method can consistently estimate the spectral density, even though the noise is no longer a restrictive independent and identically distributed (i.i.d.) process. Secondly, this promising finding is established with much less restrictive constraints than previously imposed.

E0167: **Hysteretic multivariate Bayesian structural GARCH model with soft information**
Presenter: **Shih-Feng Huang**, National Central University, Taiwan

A hysteretic multivariate Bayesian structural GARCH model with soft information, denoted by SH-MBS-GARCH, is proposed to describe multidimensional financial time-series dynamics. First, the GARCH effects inherent is filtered in each financial time series by the De-GARCH technique. Next, a hysteretic multivariate Bayesian structural model is established for the multidimensional De-GARCH time series to simultaneously capture the trend, seasonal, cyclic, and endogenous (or exogenous) covariates effects. In particular, soft information is extracted from the daily financial news and added to the model's hysteretic part to reflect economic effects on the time-series behaviour. An MCMC algorithm is proposed for parameter estimation. The empirical study employs the Dow Jones Industrial, Nasdaq, and Philadelphia Semiconductor indices from January 2016 to December 2020 to investigate the performances of the proposed model. Numerical results reveal that the SH-MBS-GARCH model has better fitting and prediction performances than competitors.

EO119 Room 703 NEW DIRECTIONS IN TIME SERIES ANALYSIS Chair: Sumanta Basu

E0817: **Exploring financial networks using quantile regression and Granger causality**
Presenter: **Samriddha Lahiry**, Harvard University, United States
Co-authors: Sumanta Basu, Diganta Mukherjee, Kara Karpman

Granger causality-based techniques to build networks among financial firms using time series of their stock returns have received significant attention recently. Existing Granger causality network methods model conditional means and do not distinguish between connectivity in the lower and upper tails of the return distribution - an aspect crucial for systemic risk analysis. Statistical methods are proposed that measure connectivity in the networks using tail-based analysis. They are able to distinguish between connectivity in the lower and upper tails of the return distribution. This is achieved using bivariate and multivariate Granger causality analysis based on regular, and Lasso penalized quantile regressions, an approach called quantile Granger causality. An asymptotic theory of quantile Granger causality estimators is provided under a quantile vector autoregressive model, showing its benefit over regular Granger causality analysis on simulated data. The proposed method is applied to the monthly stock returns of large U.S. firms and demonstrates that lower tail-based networks can detect systemically risky periods with higher accuracy than mean-based networks. In a similar analysis of large Indian banks, it is found that upper and lower tail networks convey different information about periods of high connectivity that are governed by positive vs negative news in the market.

E0878: **Asymptotic of large autocovariance matrices**
Presenter: **Monika Bhattacharjee**, IIT Bombay, India

The high dimensional moving average process is considered, and the asymptotics for eigenvalues of its sample autocovariance matrices are explored. It is proved that under quite weak conditions, in a unified way, the limiting spectral distribution (LSD) of any symmetric polynomial in the sample autocovariance matrices, after suitable centring and scaling, exists and is non-degenerate. Methods from free probability in conjunction with the method of moments to establish our results are used. In addition, a general description of the limits in terms of some freely independent variables is provided. Asymptotic normality results are shown for the traces of these matrices. Statistical uses of these results in problems such as order determination of high dimensional MA and AR processes and testing hypotheses for such processes' coefficient matrices are suggested.

E0753: **Using t-SNE in analyzing multivariate time series data**
Presenter: **Soudeep Deb**, Indian Institute of Management Bangalore, India

In multivariate data, t-distributed stochastic neighbour embedding (t-SNE) is one of the advanced dimension reduction techniques. Albeit it is primarily used for visualization in lower dimensions, in the context of time series analysis, it has the potential to be utilized in other problems too. Two different applications of t-SNE in analyzing multivariate time series are discussed. First, a classification technique that uses the advantages of dimension reduction through t-SNE is proposed, coupled with the attractive properties of nonparametric spectral density estimates and the k-nearest neighbour technique. The theoretical consistency of the proposed algorithm is proved, and the efficacy of the method is shown using an interesting dataset from medical research. The second part is about a new methodology to detect structural breaks in multivariate time series. Once again, the same principles are used, and t-SNE with nonparametric spectral density estimates in lower dimensions is utilized. Relevant empirical justifications are obtained to demonstrate the accuracy of the proposed method in detecting structural breaks in multivariate series. For application, the exchange rates of the Indian Rupee against four other major currencies from the last decade are considered.

E0898: **High-dimensional latent Gaussian count time series**
Presenter: **Marie Duker**, Cornell University, United States

The focus is on on stationary vector count time series models defined via deterministic functions of a latent stationary vector Gaussian series. The construction is very general and ensures a pre-specified marginal distribution for the counts in each dimension, depending on unknown parameters that can be marginally estimated. The Gaussian vector series injects flexibility in the model's temporal and cross-sectional dependencies, perhaps through a parametric model akin to a vector autoregression. It is discussed how the latent Gaussian model can be estimated by relating the covariances of the observed counts and the latent Gaussian series. In a possibly high-dimensional setting, concentration bounds are established for the differences between the estimated and true latent Gaussian autocovariance for the observed count series and the estimated marginal parameters. The result is applied to the case when the latent Gaussian series follows a VAR model, and its parameters are estimated sparsely through a LASSO-type procedure.

E0830: **Ordinal pattern based time series analysis**
Presenter: **Annika Betken**, University of Twente, Netherlands
Co-authors: Herold Dehling, Alexander Schnurr, Ines Nuessgen, Jeannette Woerner, Jannis Buchsteiner, Annika Betken

In time series analysis, ordinal patterns describe the relative position of consecutive realizations generated by a stochastic process. Among other things, estimators for the probabilities of occurrence of ordinal patterns (ordinal pattern probabilities) in time series are considered. Statistical properties of these estimators in discrete-time Gaussian processes with stationary increments are investigated. By means of Rao-Blackwellization, further the estimation of ordinal pattern probabilities is improved. Moreover, limit theorems that describe the asymptotic distribution of the considered estimators are established. The limit distributions may differ depending on the behaviour of the data-generating processes' autocorrelation function. As an application, the Zero-Crossing estimator is discussed for the Hurst parameter characterizing fractional Brownian motions.

EO136 Room 705 RECENT ADVANCES OF HIGH-DIMENSIONAL DATA ANALYSIS Chair: Jin-Ting Zhang

E0201: **A further study on Chen-Qin's test for two-sample Behrens-Fisher problems for high-dimensional data**
Presenter: **Tianming Zhu**, National Institute of Education, Nanyang Technological University, Singapore
Co-authors: Jin-Ting Zhang

A further study on Chen-Qin's test, namely CQ-test, for two-sample Behrens-Fisher problems for high-dimensional data is conducted, resulting in a new normal-reference test where the null distribution of the CQ-test statistic is approximated with that of a chi-squared-type mixture, which is obtained from the CQ-test statistic when the null hypothesis holds and when the two samples are normally distributed. The distribution of the chi-squared-type mixture can be well approximated by a three-cumulant matched chi-squared approximation with the approximation parameters consistently estimated from the data. The asymptotical power of the new normal-reference test under a local alternative is established. Two simulation studies demonstrate that in terms of size control, the new normal-reference test with the three-cumulant matched chi-squared-approximation performs well regardless of whether the data are nearly uncorrelated, moderately correlated, or highly correlated, and it performs substantially better than the CQ-test. A real data example illustrates the new normal-reference test.

E0243: **A fast and accurate kernel-based independence test**
Presenter: **Jin-Ting Zhang**, National University of Singapore, Singapore
Co-authors: Jin-Ting Zhang, Tianming Zhu

Testing the dependency between two random variables is a vital statistical inference problem since many statistical procedures rely on the assumption that the two samples are independent. A so-called HSIC (Hilbert-Schmidt Independence Criterion)-based test has been proposed to test whether two samples are independent. Its null distribution is approximated either by permutation or a Gamma approximation. Unfortunately, the permutation-based test is very time-consuming, and the Gamma-approximation-based test does not work well for high-dimensional data. A new HSIC-based test is proposed. Its asymptotic null and alternative distributions are established. It is shown that the proposed test is root-n consistent. A three-cumulant matched chi-squared approximation is adopted to approximate the null distribution of the test statistic. The proposed test can be applied to many different data types, including multivariate, high-dimensional, and functional data, by choosing a proper reproducing kernel. Three simulation studies and two real data applications show that the proposed test outperforms several tests for multivariate, high-dimensional, and functional data in terms of level accuracy, power, and computational cost.

E0273: **Catalytic priors: Using synthetic data to specify prior distributions in Bayesian analysis**
Presenter: **Dongming Huang**, National University of Singapore, China
Co-authors: Feicheng Wang, Donald Rubin, Samuel Kou

Catalytic prior distributions provide general, easy-to-use, and interpretable specifications of prior distributions for Bayesian analysis. They are particularly beneficial when observed data are inadequate to well-estimate a complex target model. A prior catalytic distribution stabilizes a high-dimensional "working model" by shrinking it toward a "simplified model". The shrinkage is achieved by supplementing the observed data with a small amount of "synthetic data" generated from a predictive distribution under the simpler model. This framework is applied to generalized linear models, where various strategies are proposed for specifying a tuning parameter governing the degree of shrinkage and resultant properties are studied. The catalytic priors have simple interpretations and are easy to formulate. In our numerical experiments and a real-world study, the performance of the inference based on the catalytic prior is superior to or comparable to that of other commonly used prior distributions.

E0676: **Maximum profile binomial likelihood estimation for the semiparametric Box–Cox power transformation model**
Presenter: **Tao Yu**, National University of Singapore, Singapore
Co-authors: Pengfei Li, Baojiang Chen, Jing Qin

The Box-Cox transformation model has been widely applied for many years. The parametric version of this model assumes that the random error follows a parametric distribution, say the normal distribution, and estimates the model parameters using the maximum likelihood method. The semiparametric version assumes that the random error distribution is completely unknown; existing methods either need strong assumptions or are less effective when the random error distribution significantly deviates from the normal distribution. The semiparametric assumption is adopted, and a maximum profile binomial likelihood method is proposed. The joint distribution of the estimators of the model parameters is theoretically established. Through extensive numerical studies, it is demonstrated that this method has an advantage over existing methods when the distribution of the random error deviates from the normal distribution. Furthermore, the method's performance is compared to existing methods on an HIV data set.

EO179 Room 708 SNAPSHOT ON CURRENT FUNCTIONAL DATA METHODOLOGIES Chair: Frederic Ferraty

E0210: **Fast generalized functional principal components analysis**
Presenter: **Julia Wrobel**, Colorado School of Public Health, United States

A new fast generalized functional principal components analysis (fast-GFPCA) algorithm is proposed for dimension reduction of non-Gaussian functional data. The method consists of: (1) binning the data within the functional domain; (2) fitting local random intercept generalized linear mixed models in every bin to obtain the initial estimates of the person-specific functional linear predictors; (3) using fast functional principal component analysis to smooth the linear predictors and obtain their eigenfunctions; and (4) estimating the global model conditional on the eigenfunctions of the linear predictors. An extensive simulation study shows that fast-GFPCA performs as well or better than existing state-of-the-art approaches; it is orders of magnitude faster than existing general-purpose GFPCA methods and scales up well with the number of observed curves and observations per curve. Methods were motivated by and applied to a study of active/inactive physical activity profiles obtained from wearable accelerometers in the NHANES 2011-2014 study. The method can be implemented by any user familiar with mixed model software, though the R package fastGFPCA is provided for convenience.

E0499: **Covariate adjusted mixed membership models for functional data**
Presenter: **Donatello Telesca**, UCLA, United States

Mixed membership modelling in the context of functional data analysis is discussed. The aim is to propose to leverage the multivariate KL representation of a stochastic process to induce a probabilistic representation of mixed membership to pure membership processes. In this context, covariate adjustment is discussed about both the mean and covariance functions. The motivation comes from applications in functional brain imaging through electroencephalography.

E0580: **Geometric and topological perspectives on unsupervised functional data analysis**
Presenter: **Fabian Scheipl**, Ludwig-Maximilians-Universitaet Muenchen, Germany
Co-authors: Moritz Hermann

Clustering, as well as outlier or anomaly detection, are important unsupervised tasks in functional data analysis. The problem from geometrical and topological perspectives is discussed, and a framework is provided for unsupervised FDA that exploits a functional data set's (metric) structure. The approach rests on the manifold assumption, i.e., that the observed, nominally infinite-dimensional functional data lie on or close to a much lower dimensional manifold and that this intrinsic structure can be inferred with manifold learning methods. It is shown that exploiting this structure can significantly improve the detection of outlying functions and provides a simple, robust and easily customizable way to apply well-established and highly-performant modern clustering algorithms to functional data. The framing also suggests a novel, precise, and widely applicable distinction between distributional and structural outliers based on the geometry and topology of the data manifold that clarifies conceptual ambiguities prevalent throughout the literature.

E0724: **Functional predictor selection and its nonasymptotic behavior**
Presenter: **Jun Song**, Korea University, Korea, South

A new method of functional prediction and estimation is presented in a scalar-on-function regression model. In particular, a functional adaptive group-lasso type penalization is applied to the regression problem when multivariate functional data are predictors of the functional regression model. Introducing a new penalty specific to infinite-dimensional functional data can relieve stringent assumptions for theoretical verification. The consistency of the method is shown, and the nonasymptotic behaviour of the method is investigated under a finite sample based on this new penalty type. Lastly, simulation and real data application show the effectiveness and validity of the method.

EO151 Room 709 RECENT ADVANCES IN PANEL DATA ECONOMETRICS Chair: Takahide Yanagi

E0217: **Low-rank panel quantile regression: Estimation and inference**
Presenter: **Yiren Wang**, Singapore Management University, Singapore

A class of low-rank panel quantile regression models is proposed, allowing for unobserved slope heterogeneity over individuals and time. The heterogeneous intercept and slope matrices are estimated via nuclear norm regularization followed by sample splitting, row- and column-wise quantile regressions and debasing. It is shown that the estimators of the factors and factor loadings associated with the slope matrices are asymptotically normally distributed. In addition, two specification tests are developed: one for the null hypothesis that the slope coefficient is a constant over time and/or individuals under the case that the true rank of the slope matrix equals one, and the other for the null hypothesis that the slope coefficient exhibits an additive structure under the case that the true rank of slope matrix equals two. The finite sample performance of estimation and inference via Monte Carlo simulations and real datasets are illustrated.

E0179: **What do we get from two-way fixed effects regressions? Implications from numerical equivalence**
Presenter: **Shoya Ishimaru**, Hitotsubashi University, Japan

In any multiperiod panel, a two-way fixed effects (TWFE) regression is numerically equivalent to a first-difference (FD) regression that pools all possible between-period gaps. Building on this observation, numerical and causal interpretations of the TWFE coefficient are developed. At the sample level, the TWFE coefficient is a weighted average of FD coefficients with different between-period gaps. This decomposition is useful for assessing the source of identifying variation for the TWFE coefficient. A causal interpretation of the TWFE coefficient at the population level requires a common trend assumption for any between-period gap. The assumption has to be conditional on changes in time-varying covariates. It is shown that these requirements can be naturally relaxed by modifying the estimator using a pooled FD regression.

E0478: **Using extreme bounds analysis to assess reproducibility**
Presenter: **Andrew Adrian Pua**, Xiamen University, China
Co-authors: Markus Fritsch, Joachim Schnurbus

Suppose we are willing to accept a more tolerable standard for reproducibility. Suppose we find that we cannot exactly reproduce numerical findings, but are willing to give the maximum benefit of the doubt to a published result. We show how to extend and apply extreme bounds analysis to provide a quick check of whether a published result involving instrumental variable and generalized method of moments estimates may tentatively be accepted. The minimal ingredients we need for the check are (1) output from an instrumental variables regression of the outcome variable on the regressors and the excluded instruments and (2) the cleaned data along with the published coefficient of interest and its standard error. Extending and applying extreme bounds analysis in this manner is compatible with the spirit of existing tools such as the GRIM test or statcheck.

E0196: **Fixed-T estimation of matrix-valued factor models**
Presenter: **Ying Lun Cheung**, Capital University of Economics and Business, China

Many data sets are best represented as time series of matrices, yet econometric models tailored for this data structure remain scant. The matrix-valued factor model (MVFM) is one of the few such models. Most existing methods operate under the "large N, large T" context as a natural extension of the high-dimensional factor model. However, many matrix-variate data sets have either a short time span or a low frequency. The estimation of the MVFM under the "large N, fixed T" setting is considered. It is shown that the 2DSVD procedure continues to work. The consistency and asymptotic normality of the estimator is proved. The estimator's performance is evaluated through simulations and applications with real data.

EC303 Room 603 EMPIRICAL FINANCE Chair: Feiyu Jiang

E1122: **Simulation of high-fidelity limit order book data with machine learning model for Asia exchange markets**
Presenter: **Chun Fai Carlin Chu**, The Hang Seng University of Hong Kong, Hong Kong

The aim is to address the properties of level 2 market data and discuss several workable configurations for simulating high fidelity limit order book data using Agent-Based Interactive Discrete Event Simulation (ABIDES) and world agent conditional generative adversarial network (CGAN) for Asia exchange markets with a lunch break period. The lunch break period occurs in specific exchanges in Asia regions, including Hong Kong, Shanghai, Shenzhen and Tokyo. Its existence signifies a unique characteristic which is not observed in other world exchanges, and the corresponding simulation strategy should differ. Properties of historical price value and market depth before and after a lunch break are examined, and proper ways for fine-tuning simulation model parameters are demonstrated. Furthermore, the relationships between the granularity of the condition and the fitness of the simulated data are investigated. The simulated results are evaluated under statistical and practical criteria, including MSE, non-parametric hypothesis test and VWAP benchmark.

E1173: **Do investors compensate for unsustainable consumption using sustainable assets?**
Presenter: **Emily Kormanyos**, Deutsche Bundesbank, Germany

To understand retail investor demand for sustainable assets, carbon footprints are estimated for 6,151 investors by linking their administrative consumption data to product-level carbon intensities. It is shown that compensation motives drive investments in low-emission assets and formally rule out alternative explanations. A survey with 3,646 participants reveals that investors who believe they have above-average footprints choose sustainable investing specifically as a form of compensation and significantly more often than others. Further evidence is provided that portfolio sustainability is related to religious beliefs, historically tied to offsetting and that income or sample selection effects are not driving my results.

E1207: **A similarity-based approach to covariance forecasting**
Presenter: **Mark Jennings**, University of Oxford, United Kingdom
Co-authors: Chao Zhang, Alvaro Cartea, Mihai Cucuringu

Forecasting covariance matrices of time series is a ubiquitous problem in finance. A framework which leverages recurrent structures in data to tackle the problem is introduced. The framework calculates similarity scores between test inputs and training inputs using their recent histories and uses these scores to filter the training data. Data that do not share the dynamics of the test input are excluded from the regression, reducing the complexity of the forecasting task. Then forecasts based only on the relevant training data are produced using simple non-parametric and linear methods. Furthermore, a dynamic empirical similarity ensemble scheme is proposed, which generates a weighted average of individual forecasts based on recent performance. The framework produces computationally efficient and interpretable forecasts of the realised covariances of US equity returns and outperforms widely-used benchmark models. The framework adjusts to rapidly changing market conditions by down-weighting models that are unsuitable for the current market dynamics, which minimises the impact of model specification and leads to improved robustness against turbulent conditions compared to alternative models. The economic value of the forecasts is also evaluated by applying them to portfolio optimisation, and it is shown that the framework generates a higher Sharpe ratio than those of competing models.

E1220: **Can price jumps be explained by leverage effect of volatility? A HAR-Hawkes framework**
Presenter: **Ping Chen Tsai**, National Sun Yat-sen University, Taiwan

The predicting nature of price jumps in a HAR-Hawkes framework is investigated. Using four equity market indices of 5-minute returns up to 10 years, First, it is shown that by adding persistent leverage effect in the HAR-RV model, many extreme out-of-sample residuals standardized by the conditional expectation of Realized Volatility - called pseudo jumps - can be explained away. Then two jump tests (forward- and backwards-looking) with two intraday volatility pattern estimators (conventional and jump-robust) are considered; a large portion of price jumps are contained in the set of pseudo jumps, and the persistent leverage effect explains some of these price jumps. This result is more prominent when a forward-looking test gives price jumps. The Hawkes process is further applied, showing that the conditional intensity of price jumps behaves differently than the pseudo jumps. The results suggest that price jumps are self-exciting and can also be predicted by the leverage effect of volatility.

EC300 Room 605 MACROECONOMETRICS Chair: Caleb Miles

E0444: **Estimating the interest rate trend in a shadow rate term structure model**
Presenter: **Jun Ma**, Northeastern University, United States
Co-authors: Yang Han

A shadow rate no-arbitrage dynamic term structure model with drifting trends is proposed to estimate the long-run trend of real interest rates using data from the U.S., the U.K., and Germany from January 1972 to April/March 2022. The interest rate trends of all three countries have declined since the 1990s, and there is strong co-movement among them. It is documented evidence that this model can provide better yield forecasts than existing models. The term premium estimates of the model are stationary and are positively correlated with inflation uncertainty measures.

E0654: **Improving output gap estimation: A bottom-up approach**
Presenter: **Sina Streicher**, ETH Zurich, KOF Swiss Economic Institute, Switzerland
Co-authors: Alexander Rathke

A multivariate Bayesian state space model is proposed to identify potential output and the output gap consistent with the dynamics of the underlying production sectors of the economy and those of inflation and the labour market. This approach allows decomposing economic fluctuations and long-term trend growth in output and employment into its driving factors. Tracking the cycles of individual sectors allows policy actions to be targeted at specific industries, thereby increasing their efficiency and reducing the chance of pro-cyclical outcomes. Applying the model to the Swiss economy reveals substantial divergence among the considered production sectors. The Swiss economy's business cycle and growth potential are most clearly influenced by the sectors most dependent on the global economy - manufacturing, financial, and other economic services. Compared to established estimation approaches, the enriched information set can help paint a more comprehensive picture of the business cycle.

E0852: **Macroprudential stress testing: A proposal for the United States fund sector**
Presenter: **Bui Dieu Thao Nguyen**, Le Mans University, France
Co-authors: Thi Thanh Huyen Nguyen

The first part assesses the aggregate vulnerability of the US fund sector by implementing an empirical framework for macro-prudential stress testing. First, t investors' behaviour is captured in response to adverse funds performance by assuming the non-linearity of the flow performance sensitivity (FPS) in the Markov Switching VAR. Second, the second-round effect comprising impacts of additional funding shocks and asset fire sales on the funds' resilience is accounted for by estimating the time-varying price impact ratio (based on the Amihud ratio). The model finally defines an indicator of aggregate vulnerability for each US fund category. Empirical results showed that limited degrees of vulnerability were found for almost all categories of funds. The growth funds have been the most vulnerable, followed by large C funds and MicroC funds. This implies that the investment fund sector in the US does not raise any particular concerns about financial stability as of September 2018. The second part assesses the interconnectedness in the US fund sector. A VAR estimation is performed on a system of aggregate vulnerability indicators of the six categories of funds. Then the variance decompositions are applied to detect how much of the future uncertainty associated with the stress in fund category i can be explained by stress shocks in fund category j. Empirical results showed that funds exposed to less liquid asset classes are more likely to be affected by shocks from others.

E1176: **The speed of state-level recoveries**
Presenter: **Luiggi Donayre**, University of Minnesota - Duluth, United States
Co-authors: Irina Panovska

The abrupt decline in U.S. GDP in 2020:Q2 has reignited an interest in the shape of recessions. Aggregate models assume that business cycles are similar across states. However, several economic ideas suggest that recessions and their recoveries may be idiosyncratic to each state. For example, those triggered by financial crises could generate longer-lasting effects, indicating L-shaped recessions. Meanwhile, downward wage rigidity may generate contractions followed by recoveries of similar amplitude, implying U-shaped recessions. Since financial crises and labour market features are heterogeneous, the speed of recovery across states, and thus recession shapes, may vary widely. To study these issues, a Markov-switching model is estimated, augmented with a bounce-back effect, to U.S. state-level output growth and find large differences across states. At the aggregate level, the bounce-back parameter 0.32 suggests relatively slow recoveries. State-level bounce-back estimates range from 0.05 to 1.36, evidencing large differences in the speed of recovery. Southern and rust-belt states exhibit slow recoveries, suggesting more L-shaped recessions. Meanwhile, mountain and oil-producing states exhibit faster, U-shaped recoveries. Because economic policies are designed to smooth aggregate business cycle fluctuations, such policies' effects vary widely across states and regions.

EC289 Room 701 APPLIED STATISTICS Chair: Wendun Wang

E0307: **Censored experiments for computing the average run length**
Presenter: **Sungim Lee**, Dankook University, Korea, South
Co-authors: Johan Lim

The purpose is to introduce a simple and efficient method for computing the average run length, commonly used to measure control chart performance. Generally, a large positive number is assumed, and then many run lengths are taken to compute the average run length in a simulation which is very time-consuming. Moreover, deleting cases with a run length larger than the predetermined maximum necessarily causes bias in computing the average run length. The method suggests this step is unnecessary if the predetermined run length can be represented by type I censored data. Assuming memoryless run lengths, the mean and standard deviation are estimated. Traditional Monte Carlo simulation is compared with the proposed procedure, including Markov chain approximation or integration methods, depending on the control chart type (Shewhart-type, EWMA, or CUSUM). The results show that the proposed methods outperform traditional methods, demonstrating the approach's applicability across various scenarios.

E0810: **Comparing product quality using a distribution-wise index**
Presenter: **Tsai-Yu Lin**, Feng Chia University, Taiwan

During the production process, producers often find it necessary to compare quality characteristic means to confirm and verify whether their improvement processes are effective. However, these methods probably ignore the individual variations between the samples. The concept of distribution-wise comparison criterion (DCC) is proposed instead of the traditional approach. The procedures of DCC for specific distributions are illustrated based on the conformance proportion in two examples. For given the reference group of the conformance proportion, the specification limits are calculated, then use these specification limits to compute the conformance proportion of the test group and then estimate a confidence lower bound for the conformance proportion of the test group by using the concept of fiducial generalized pivotal quantity. The proposed procedure has proven useful through detailed simulation results.

E1018: **The mean group estimators for multi-level autoregressive models with intensive longitudinal data**
Presenter: **Boyan Yin**, Hiroshima University, China
Co-authors: Kazuhiko Hayakawa

The mean group (MG) estimators are proposed to estimate multilevel (vector) autoregressive models with intensive longitudinal data. The MG estimator is originally proposed in econometrics, but is new to the behavioural science. Since the naive MG estimator suffers from the small sample bias problem, jackknife and analytical bias corrections are proposed. It is argued that the MG estimator has several advantages over existing methods, such as restricted maximum likelihood or Bayesian methods in terms of model specification and implementation. Monte Carlo simulation is performed to investigate the performance of the MG estimators and compare them with the existing methods. The simulation results indicate that the bias-corrected MG estimators have superior or comparable performance compared to the existing methods.

E1132: **Gaussian mixture models for changepoint detection**
Presenter: **Utkarsh Dang**, Carleton University, Canada
Co-authors: Wangshu Tu, Sanjeena Dang

Changepoint detection aims to find abrupt changes in time series data. These changes denote substantial modifications to the process; they can vary from simple changes in location to a change in distribution. Traditional changepoint detection methods often rely on a cost function to assess if a change occurred in a series. Here, changepoint detection in a clustering framework is investigated, and a novel changepoint detection algorithm is developed using a finite mixture of regressions with concomitant variables. Through the introduction of a label correction mechanism, the unstructured cluster labels are treated as ordered and distinct segment labels. Different kinds of change can be captured using a parsimonious family of models. Performance is illustrated on simulated and real data.

EC271 Room 702 NETWORKS AND GRAPHICAL MODELS Chair: Boris Choy

E1001: **Assessing weather risk: A non-parametric test for network independence with distance covariance**
Presenter: **Pok Him Cheng**, The Chinese University of Hong Kong, Hong Kong

Network models have received increasing popularity in recent years because of the growing availability of large-scale social network data and the need to model complex systems such as meteorology and biology. First, the distance covariance of a set of random vectors in a network is discussed by extending the existing work on distance covariance, where the latter has the crucial feature that it equals zero if and only if two random vectors are independent and thus can detect arbitrary types of non-linear associations. The proposed measure includes special cases such as the auto-distance covariance in time series and random fields. Based on the new measure, a new test for the independence of network data is developed. In particular, a Ljung-Box-type test for associative autocorrelation in a graph-structured network setting is proposed. Extensive simulation studies with various dependency structures illustrate the test's usefulness. The proposed method often outperforms many prevalent ones in the literature, especially when the data exhibits a non-linear relationship. The asymptotic distributions of the test statistics are established under different network structures with the aid of incomplete U-Statistics. The test is applied to study the goodness-of-fit of a fitted network model based on the residuals. An example is demonstrated using England and Wales's climate wind speed data, fitted by a generalized network autoregressive model with spatial and temporal components.

E1163: **Structure learning of graphical models for count data, with applications to single-cell RNA sequencing**
Presenter: **Thi Kim Hue Nguyen**, University of Padova, Italy
Co-authors: Davide Risso, Monica Chiogna, Koen Van Den Berge

The problem of estimating the structure of a graph from observed data is of growing interest in the context of high-throughput genomic data and single-cell RNA sequencing in particular. These, however, are challenging applications since the data consist of high-dimensional counts with high variance and over-abundance of zeros. Here, general frameworks for learning the structure of a graph from single-cell RNA-seq data are presented. It is demonstrated with simulations that the approaches can retrieve the structure of a graph in a variety of settings, and it is shown the utility of the approach on real data.

E1313: **Two-sample test for stochastic block models via maximum entry-wise deviation**
Presenter: **Kang Fu**, Central China Normal University, China
Co-authors: Jianwei Hu

The stochastic block model is a popular tool for detecting community structures in network data. Detecting the difference between two community structures is an important issue for stochastic block models. However, the two-sample test has been a largely under-explored domain, and too little work has been devoted to it. Based on the maximum entry-wise deviation of the two centred and rescaled adjacency matrices, a novel test statistic is proposed to test two samples of stochastic block models. The null distribution of the proposed test statistic is proved to converge in distribution to a Gumbel distribution, and the change of the two samples from stochastic block models can be tested via the proposed method. Then, it is shown that the proposed test has an asymptotic power guarantee against alternative models. One noticeable advantage of the proposed test statistic is that the number of communities can be allowed to grow linearly up to a logarithmic factor. Further, the proposed method is also extended to the degree-corrected stochastic block model. Both simulation studies and real-world data examples indicate that the proposed method works well.

E1214: **Extension of LiNGAM to functional data**
Presenter: **Tianle Yang**, Osaka University, Japan
Co-authors: Joe Suzuki

A causal order is considered, such as the cause and effect among variables. In the Linear Non-Gaussian Acyclic Model (LiNGAM), the order can only be identified if at least one of the variables is non-Gaussian. The notion of variables is extended to functions (Functional Linear Non-Gaussian Acyclic Model, Func-LiNGAM). First, it is proved that the order among random functions, if one of them is a non-Gaussian process, can be identified. In the actual procedure, the functions are approximated by random vectors. To improve the correctness and efficiency, optimising the coordinates of the vectors in such a way as functional principal component analysis is proposed. The experiments contain an order identification simulation among multiple functions for given samples. In particular, the Func-LiNGAM is applied to recognize the brain connectivity pattern with fMRI data. Improvements in accuracy and execution speed compared to existing methods can be seen.

EC260 Room 704 COMPUTATIONAL STATISTICS AND ECONOMETRICS Chair: David Nott

E0728: **A minibatch Gibbs sampler for scalable large-scale Bayesian inference on latent variable models**
Presenter: **Dongrong Li**, The Chinese University of Hong Kong, Hong Kong

Efficient and scalable Markov Chain Monte Carlo (MCMC) algorithms are essential for modern Bayesian computation since evaluating the joint density on the full dataset in each iteration is computationally prohibitive in the big-data era. Mini batching has emerged as a strategy to tackle this problem, but existing methods require non-trivial upper or lower bounds of the joint density or heavily rely on gradient evaluation. A novel minibatch Gibbs sampler for large-scale Bayesian latent variable models is proposed, which can efficiently sample from an approximate posterior density. The sampler uses the variable splitting technique and introduces auxiliary variables to ensure efficient mini-batching at each iteration. It's flexible, can handle any Metropolis proposal, and doesn't require non-trivial upper or lower bounds of the joint density. It is shown that the approximate density can be arbitrarily close to the true posterior asymptotically and establish explicit non-asymptotic error and mixing bounds to theoretically guarantee the convergence rates. The sampler's performance on synthetic and real data is evaluated, demonstrating its advantages over existing algorithms. The proposed minibatch Gibbs sampler offers a flexible, efficient, and scalable solution for large-scale Bayesian latent variable models and has the potential to advance modern Bayesian computation.

E1120: **Generalized linear models for massive data via doubly-sketching**
Presenter: **Jason Hou-Liu**, University of Waterloo, Canada
Co-authors: Ryan Browne

Generalized linear models are a popular analytics tool with interpretable results and broad applicability but require iterative estimation procedures that impose data transfer and computational costs that can be problematic under some infrastructure constraints. A doubly-sketched approximation of the iteratively re-weighted least squares algorithm is proposed to estimate generalized linear model parameters using a sequence of surrogate datasets. The procedure repeatedly sketches to both reduce data transfer costs and reduce data computation costs, yielding wall-clock time savings in approximating the regression coefficients and standard errors. Asymptotic properties of the proposed procedure are shown, with empirical results from simulated and real-world datasets. The efficacy of the proposed method is investigated across a variety of commodity computational infrastructure configurations accessible to practitioners. A highlight is the estimation of a Poisson-log generalized linear model across almost 1.7 billion observations on a personal computer in 25 minutes.

E1182: **Maximum contribution to the likelihood: An estimation approach for stochastic expectation-maximization algorithm**
Presenter: **Alexander Sharp**, University of Waterloo, Canada
Co-authors: Ryan Browne

The stochastic EM algorithm replaces the E-step with a Monte Carlo approximation, trading monotonicity for the potential to escape local maxima. Estimation techniques include averaging the tail of the chain and choosing the value in the chain associated with the largest likelihood value. It is demonstrated that the latter estimator diverges from the maximum likelihood estimate with high probability as the dimensionality of the parameter increases but that it is also more precise in terms of chain length when the parameter is a scalar. Based on these findings, a new estimator is proposed which achieves this same level of precision for the inference of multidimensional parameters is proved. Simulation studies demonstrate the benefits of the proposed estimator when compared to topical approaches.

E1282: **Efficient algorithms for large-scale optimal transport problems**
Presenter: **Cheng Meng**, Renmin University of China, China

Optimal transport methods have become increasingly predominant in machine learning, deep learning, statistics, computer vision, and biomedical research. Despite the wide application, existing methods for solving optimal transport problems may suffer from a substantial computational burden when the sample size is large. The Spar-Sink algorithm is introduced to alleviate the computational burden, which utilizes a novel importance sparsification scheme to accelerate the popular Sinkhorn algorithm. This approach can be effectively applied to the entropic optimal transport problem, unbalanced optimal transport problem, Gromov-Wasserstein distance approximation, Wasserstein barycenter estimation, and generative modelling, among others. The application of echocardiography is considered, which is one of the most promising techniques to display the movements of the myocardium. Experiments demonstrate the approach can effectively identify and visualize cardiac cycles, from which one can diagnose heart failure and arrhythmia. To evaluate the numerical accuracy of cardiac cycle prediction, the task of predicting the end-diastole time point using the end-systole one is considered. Results show Spar-Sink performs as well as the classical Sinkhorn algorithm, requiring significantly less computational time.

Wednesday 02.08.2023 16:10 - 17:50 Parallel Session I – EcoSta2023

EO200 Room 02 ADVANCES IN TIME SERIES ECONOMETRICS Chair: Jihyun Kim

E0207: **Instrumental factor models for high-dimensional functional data**
Presenter: **Young-Kwang Kim**, Toulouse School of Economics, Korea, South
Co-authors: Jihyun Kim

The instrumental factor model is introduced that extends conventional factor models in two directions. First, the factor model is developed for high-dimensional data, from scalar-valued data to functional data that has gained fast-growing popularity. Second, it is well-known that the conventional estimation approach based on the principal component analysis (PCA) requires both cross-sectional dimension and the time horizon of data to be large. Under the proposed approach that utilizes additional characteristic variables as instruments, the estimators achieve consistency as long as the cross-sectional dimension is large enough. The eigenuses value ratio method is then proved consistently aided to estimate the unknown number consistently. The numerical experiments confirm that the estimation approach outperforms the conventional PCA-based method, especially for short panel da a. In concussion, with an empirical application to analyze the long-run relationship between global warming and world GDP. The results support the growing consensus that human activities are the dominant cause of global warming.

E0212: **A trajectories-based approach to measuring intergenerational mobility**
Presenter: **Seunghee Lee**, Korea Development Institute, Korea, South

An approach is developed to intergenerational mobility in which the trajectories of childhood and adolescent family characteristics define the conditioning objects for characterizing mobility across generations. This contrasts with standard approaches in which family influences are summarized by scalar measures such as permanent income. This perspective leads to functional regression methods that measure how parental incomes and family structures at different points in time are associated with future outcomes. Collections of trajectories that lead to relative success versus deprivation in children have been characterized and produce novel insights into the determinants of mobility versus persistence. Offspring socioeconomic success is associated with average parental income across childhood and adolescence and with tandem trajectories. When interactions are allowed between incomes at different ages, a complex pattern of local substitution and nonlocal complementarity effects is found. Applications of the tools to offspring education and occupation produce very consistent findings to those for income.

E0213: **Inference on nonstationarity and common stochastic trends in high-dimensional or functional time series**
Presenter: **Won-Ki Seo**, University of Sydney, Australia
Co-authors: Morten Nielsen, Dakyung Seong

Statistical inference concerns unit roots and cointegration for time series taking values in a Hilbert space of an arbitrarily large, possibly infinite, and/or unknown dimension. When such a time series is given, an essential first step is to estimate the number of stochastic trends, which indicates how many linearly independent unit root processes are embedded in the time series. Statistical inference on the number of stochastic trends that remains asymptotically valid even when the time series of interest takes values in a Hilbert space of an arbitrary and indefinite dimension is developed. This has wide applicability in practice; for example, in the case of cointegrated vector time series of finite dimension, in a high-dimensional factor model that includes a finite number of nonstationary factors, in the case of cointegrated curve-valued (or function-valued) time series, and nonstationary dynamic functional factor models.

E0219: **Is there an information channel of monetary policy?**
Presenter: **Boreum Kwak**, Bank of Korea, Korea, South
Co-authors: Alexander Kriwoluzky, Oliver Holtemoeller

Three structural shocks are identified by exploiting the heteroskedasticity of the changes in short-term and long-term interest rates and exchange rates around the FOMC announcement. Studying their effects on financial variables and estimating their dynamic impact on the economy shows that two of these shocks are conventional and unconventional monetary policy shocks, respectively. The third shock leads to an increase in the stock market, industrial production, the CPI and the commodity price index. At the same time, the excess bond premium and the uncertainty index decrease, and the Dollar depreciates. It combines all the characteristics of a central bank information shock. Notably, the shock is not predictable from the news.

EO099 Room 03 HIGH DIMENSIONAL REGRESSION IN BIOMEDICAL APPLICATIONS Chair: Johan Lim

E0287: **Bayesian multi-task learning for medicine recommendation based on online patient reviews**
Presenter: **Xinlei Wang**, University of Texas at Arlington, United States
Co-authors: Yichen Cheng, Yusen Xia

A drug recommendation system that integrates information from both structured data (patient demographic information) and unstructured texts (patient reviews) is proposed. The core of the recommendation system is a multi-task learning model that predicts review ratings of several satisfaction-related measures for a given medicine, where related tasks can learn from each other when predicting. The learned models can then be applied to new patients for drug recommendation. This fundamentally differs from the widely used recommender systems in e-commerce, which do not work well for new customers (referred to as the cold-start problem). Both topic modelling and sentiment analysis are used to extract information from the review texts. The results indicate that the extracted topics help identify each drug's key features and can sometimes (but not always) help predict ratings. A variable selection component is incorporated in the model through Bayesian LASSO, which aims to filter out irrelevant features. The proposed method is effective even with a small sample size and few available features. The method is tested on two sets of drug reviews involving 17 depression or high-blood-pressure-related drugs in total, and the prediction performance is compared with existing benchmark models.

E0533: **High dimensional tests on multivariate regressions under confounding**
Presenter: **Shota Katayama**, Keio University, Japan

In high dimensional data analysis, especially in differential gene expression analysis, detecting the difference of a huge number of features between two groups is an essential problem. A unified inference on high dimensional parameters is provided for comparing two groups in the case where random assignment is not feasible. To achieve this, multivariate regressions with high dimensional response vector is considered, taking into account observable confounding variables. Based on an efficient score function, global and multiple testing procedures on its high dimensional parameter are proposed to ensure asymptotic validity in high dimensions. Applying real RNA-seq data on Covid-19 patients demonstrates significant genes involved in serious cases.

E0935: **Selection problems in multiple instance learning**
Presenter: **Seongoh Park**, Sungshin Womenś University, Korea, South
Co-authors: Johan Lim, Xinlei Wang, Tao Wang

In multiple instance, learning (MIL), a bag represents a sample that has a set of instances, each of which is described by a vector of explanatory variables, but the entire bag only has one label/response. Though many methods for MIL have been developed to date, few have paid attention to the interpretability of models and results. Two different models are considered to select instances or variables simultaneously. The first model is a Bayesian hierarchical regression model that addresses two selection problems simultaneously. To do it, the shotgun stochastic search algorithm is modified to fit in the MIL context. The model is applied to the musk data to predict binding strengths between molecules (bags) and receptors (instances). Another approach is multiple instance neural networks based on sparse attention. The sparse attention structure drops out uninformative instances in each bag, achieving both interpretability and better predictive performance. It is applied to a cancer detection problem where the one-to-many correspondence between a patient and multiple T cell receptors (TCR) sequences hinders researchers from simply adopting classical statistical/machine learning methods. Recent attempts to model this type of data still have room for improvement, especially for certain cancer types. Furthermore, explainable neural network models are not fully investigated. The proposed method aims to fill this gap.

E0996: **Models for cluster randomised designs using ranked set sampling**
Presenter: **Omer Ozturk**, The Ohio State University, United States
Co-authors: Olena Kravchuk, Richard Jarrett

Cluster randomized designs (CRD) provide a rigorous development for randomization principles for studies where treatments are allocated to cluster units rather than the individual subjects within clusters. It is known that CRDs are less efficient than completely randomized designs since the randomization of treatment allocation is applied to the cluster units. To mitigate this problem, ranked set sampling design from survey sampling studies is embedded into CRD for the selection of cluster and subsampling units. It is shown that ranking groups in ranked set sampling act like a covariate, reducing the expected to mean squared cluster error and increasing the precision of the sampling design. An optimality result is provided to determine the sample sizes at the cluster and sub-sample levels. The proposed sampling design is applied to a longitudinal study from an education intervention program.

EO025 Room 04 SURVIVAL ANALYSIS WITH MEDICAL AND HEALTH DATA SCIENCE Chair: Takeshi Emura

E0244: **A functional bivariate copula joint model for longitudinal measurements and time-to-event data**
Presenter: **Zili Zhang**, The University of Manchester, United Kingdom

A bivariate functional copula joint model, which models the repeatedly measured longitudinal outcome at each time point with the survival data, jointly by both the random effects and bivariate functional copulas, is proposed. A regular joint model normally supposes some subject-specific latent random effects or classes shared by the longitudinal and time-to-event processes. They are assumed to be conditionally independent, given these latent random variables. Under this assumption, the joint likelihood of the two processes can be easily derived, and the unobservable latent random variables naturally introduce the association between them and heterogeneity among the population. However, because of the unobservable nature of these latent variables, the conditional independence assumption is difficult to verify. Therefore, a bivariate functional copula is introduced into a regular joint model to account for the cases where there could be an extra association between the two processes which the latent random variables cannot capture. The proposed model includes a regular joint model as a particular case when the correlation function under the bivariate Gaussian copula is constant at 0. Simulation studies and dynamic predictions of survival probabilities are conducted to compare the performance of the proposed model with the regular joint model, and a real data application on the Primary biliary cirrhosis (PBC) data is performed.

E0515: **Statistical methods for integrating longitudinal and cross sectional data: A real case study**
Presenter: **Hui-Wen Lin**, Soochow University, Taiwan

Longitudinal data (including survival tracking data) is crucial in medical research as it enables researchers to detect subject changes over time. However, there are many challenges in analyzing longitudinal data, such as complicated probability functions and data gaps. A two-stage approach is proposed to collect cross-sectional and longitudinal data effectively while reducing costs. Statistical methods used for analyzing longitudinal data were conducted, and a statistical method that integrates longitudinal and cross-sectional data was proposed to correct estimation biases. This method was also used to investigate the association between chronic obstructive pulmonary disease (COPD) and herpes zoster (HZ). The results suggest that COPD may increase the risk of HZ, even after adjusting for potential confounders. The proposed method provides a way to account for missing confounders and reduce bias in observational studies.

E0844: **Parametric distributions for survival analysis, a review and historical sketch**
Presenter: **Nanami Taketomi**, Graduate school of Kurume University, Japan
Co-authors: Kazuki Yamamoto, Christophe Chesneau, Takeshi Emura

During its 330 years of history, parametric distributions have been useful for survival analysis. A lot of parametric distributions have ever been proposed and used in the medical and other fields. The historical backgrounds and statistical properties of some parametric distributions used in survival analysis are comprehensively reviewed. It is explained how the important parametric distributions have been adopted in survival analysis with original and state-of-the-art references. The exponential, Weibull, lognormal, Pareto (types I, II, and IV), and Gompertz distributions are mainly covered. These distributions may be useful to the medical data. Finally, the examples of applying the parametric distributions are shown to real datasets.

E0940: **Distributed Cox proportional hazards regression using summary-level information**
Presenter: **Dongdong Li**, Harvard Medical School, United States

Individual-level data sharing across multiple sites can be infeasible due to privacy and logistical concerns. A general distributed methodology is proposed to fit Cox proportional hazards models without sharing individual-level data in multi-site studies. Inferences are made on the log hazard ratios based on an approximated partial likelihood score function that uses only summary-level statistics. This approach can be applied to stratified and unstratified models, accommodate discrete and continuous exposure variables, and permit the adjustment of multiple covariates. In particular, stratified Cox models can be fitted with only one file transfer of summary-level information. The asymptotic properties of the proposed estimators are derived, and the proposed estimators are compared with the maximum partial likelihood estimators using pooled individual-level data and meta-analysis methods through simulation studies. The proposed method uses a real-world data set to examine the effect of sleeve gastrectomy versus Roux-en-Y gastric bypass on time to first postoperative readmission.

EO135 Room Virtual R01 BAYESIAN NON-PARAMETRIC MODELLING WITH APPLICATIONS Chair: Andrea Cremaschi

E0335: **Cox-Hawkes: Doubly stochastic spatiotemporal Poisson processes**
Presenter: **Xenia Miscouridou**, Imperial College London, United Kingdom

Hawkes processes are point process models used to capture self-excitatory behaviour in social interactions, neural activity, earthquakes and viral epidemics. They can model the occurrence of the times and locations of events. A new class of spatiotemporal Hawkes processes is developed to capture both triggering and clustering behaviour, and an efficient method is provided for performing inference. A log-Gaussian Cox process (LGCP) is used as prior for the background rate of the Hawkes process, which gives arbitrary flexibility to capture a wide range of underlying background effects (for infectious diseases, these are called endemic effects). The Hawkes process and LGCP are computationally expensive due to the former having a likelihood with quadratic complexity in the number of observations and the latter involving inversion of the precision matrix, which is cubic in observations. A novel approach is proposed to perform MCMC sampling for the Hawkes process with LGCP background, using pre-trained Gaussian Process generator,s which provide direct and cheap access to samples during inference. The efficacy and flexibility of the approach in experiments are shown on simulated data, and the methods are used to uncover the trends in a dataset of reported crimes in the US.

E1030: **Model-based clustering for categorical data via Hamming distance**
Presenter: **Raffaele Argiento**, Universita degli Studi di Bergamo, Italy
Co-authors: Lucia Paci, Edoardo Filippi-Mazzola

A model-based approach is introduced for clustering categorical data with no natural ordering. The proposed method exploits the Hamming distance to model categorical data by defining a family of probability mass functions. The elements of this family are considered kernels of a finite mixture model with an unknown number of components. Fully Bayesian inference is provided using a sampling strategy based on a trans-dimensional blocked Gibbs sampler, facilitating the computation with respect to the customary reversible-jump algorithm. Model performances are assessed via a simulation study, showing improvements in clustering recovery over existing approaches. Finally, the method is illustrated with an application to reference datasets.

E0502: **Bayesian learning of graph substructures**
Presenter: **Maria De Iorio**, National University of Singapore, Singapore
Co-authors: Willem van den Boom, Alexandros Beskos

Graphical models provide a powerful methodology for learning the conditional independence structure in multivariate data. The inference is often focused on estimating individual edges in the latent graph. Nonetheless, there is increasing interest in inferring more complex structures, such as communities, for multiple reasons, including more effective information retrieval and better interpretability. Stochastic block models offer a powerful tool to detect such structures in a network. Thus exploiting random graph theory, advances are proposed and embedding them within the graphical models' framework. A consequence of this approach is the propagation of the uncertainty in graph estimation to large-scale structure learning. Bayesian nonparametric stochastic block models as priors on the graph are considered. Such models are extended to consider clique-based blocks and multiple graph settings, introducing a novel prior process based on a Dependent Dirichlet process. Moreover, a tailored computation strategy of Bayes factors for block structure based on the Savage-Dickey ratio is devised to test for the presence of a larger structure in a graph. Our simulation approach is demonstrated in real-data applications in finance and transcriptomics.

E0925: **Joint random partition models for multivariate change point analysis**
Presenter: **Garritt Page**, Brigham Young University, United States
Co-authors: Jose Javier Quinlan Binelli, Mauricio Castro

Change point analyses identify positions of an ordered stochastic process that undergo abrupt local changes of some underlying distribution. When multiple procedures are observed, information regarding the change point positions is often shared across the different approaches. A method that takes advantage of this type of information is described. Since the number and position of change points can be described through a partition with contiguous clusters, this approach develops a joint model for these types of partitions. Computational strategies associated with our approach are described, and improved performance in detecting change points through a small simulation study is illustrated. This method is applied to a financial data set of emerging markets in Latin America, and interesting insights discovered due to the correlation between change point locations among these economies are highlighted.

EO100 Room Virtual R02 MACHINE LEARNING THEORY AND ROBUSTNESS Chair: Yiming Ying

E1039: **Convergence analysis for functional online learning algorithms**
Presenter: **Zheng-Chu Guo**, Zhejiang University, China
Co-authors: Xin Guo, Lei Shi

Convergence analysis of online stochastic gradient descent algorithms for functional linear models is reported. Adopting the characterizations of the slope function regularity, the kernel space capacity, and the sampling process covariance operator capacity, significant improvement in the convergence rates is achieved. Both prediction and estimation problems are studied, showing that capacity assumption can alleviate the convergence rate saturation as the target function's regularity increases. It is shown that with a properly selected kernel, capacity assumptions can fully compensate for the regularity assumptions for prediction problems (but not estimation problems). This demonstrates the significant difference between functional data analysis prediction and estimation problems.

E1110: **Lp-consistency of regularized kernel methods and its connection to risk consistency**
Presenter: **Hannes Koehler**, University of Bayreuth, Germany

It is well-known that risk consistency is a property that is satisfied by regularized kernel methods such as support vector machines under mild conditions. The close connection of risk consistency to L_p-consistency is investigated and established for a considerably wider class of loss functions than has been done before. From this, it is derived that the examined regularized kernel methods are indeed L_p-consistent as well. The attempt to transfer this result to shifted loss functions surprisingly reveals that this shift does not reduce the assumptions needed on the underlying probability measure to the same extent as it does for many other results regarding regularized kernel methods.

E1179: **Mathematical foundations of outcome weighted learning in precision medicine**
Presenter: **Daohong Xiang**, Zhejiang Normal University, China

Outcome-weighted learning (OWL) is one of the most popular algorithms for estimating the optimal individualized treatment rules in precision medicine. The convergence theory of OWL for the cases of bounded and unbounded clinical outcomes is mainly studied. Fast learning rates of OWL associated with least square loss, exponential-hinge loss and r-norm SVM loss are derived explicitly.

E1280: **Generalization analysis for contrastive deep representation learning**

Presenter: **Yiming Ying**, State University of New York at Albany, United States

The performance of machine learning (ML) models often depends on the representation of data, which motivates a resurgence of contrastive representation learning (CRL) to learn a representation function. Recently, CRL has shown remarkable empirical performance, and it can even surpass the performance of supervised learning models in various domains, such as computer vision and natural language processing. Recent progress is presented in establishing the learning theory foundation for CRL. In particular, the following two theoretical questions are addressed: 1) how would the generalization behaviour of downstream ML models benefit from the representation function built from positive and negative pairs? 2) Especially how would the number of negative examples affect its learning performance? Specifically, generalization bounds for contrastive learning can be shown that do not depend on the number k of negative examples up to logarithmic terms. The analysis uses structural results on empirical covering numbers and Rademacher complexities to exploit the Lipschitz continuity of loss functions. For self-bounding Lipschitz loss functions, the results are further improved by developing optimistic bounds, which imply fast rates in a low noise condition. The results are applied to learning with both linear representation and nonlinear representation by deep neural networks, for both of which explicit Rademacher complexity bounds are derived.

EO201 Room 201 TREATMENT EFFECT HETEROGENEITY AND RELATED TOPICS Chair: Seojeong Jay Lee

E0329: **Tests for heterogeneous treatment effect**

Presenter: **Fangzhou Yu**, University of New South Wales, Australia

Two hypothesis tests are developed for heterogeneous treatment effects. The method is focused on the null hypothesis that the conditional treatment effects are zero for all covariate values and the null hypothesis that the conditional treatment effects are constant for all covariate values. The tests are applied to the treatment effects identified under the unconfoundedness assumption and the local effects identified by a binary instrumental variable. The test statistics are based on the Wald statistic of the best linear projection coefficients of the treatment effects on the covariates with coefficients estimated by regressing the augmented inverse propensity-weighted outcome on the covariates. First parametric tests assuming parametric forms of the potential outcomes and the propensity score are derived, and then the parametric assumptions are relaxed by allowing for nonparametric/high-dimensional models to derive semiparametric tests where Double/Debiased Machine Learning estimates the projection coefficients. The finite sample performance of the tests is demonstrated using simulated experimental and survey datasets. The use of the tests in two applications is illustrated regarding the effect of being the only child on the mental health of the only children and the effect of 401(k) participation on the net financial assets of the participants.

E0339: **Nonparametric estimation of sponsored search auctions and impacts of ad quality on search revenue**

Presenter: **Dongwoo Kim**, Simon Fraser University, Canada

Co-authors: Pallavi Pal

The aim is to present an empirical model of sponsored search auctions in which advertisers are ranked by bid and ad quality. A new nonparametric estimator is introduced for the advertiser's ad value and its distribution under the *'incomplete information'* assumption. The ad value is characterized by a tractable analytical solution given observed auction parameters. Using Yahoo! search auction data, value distributions are estimated, and the bidding behaviour is studied across product categories. It is found that advertisers shade their bids more when facing less competition. Counterfactual analysis is also conducted to evaluate the impact of score squashing on the auctioneer's revenue. The results show that product-specific score squashing can enhance auctioneer revenue at the expense of advertiser profit and consumer welfare.

E0716: **Synthetic controls with multiple outcomes: estimating the effects of NPIs in the COVID-19 Pandemic**

Presenter: **Seojeong Jay Lee**, Seoul National University, Korea, South

Co-authors: Valentyn Panchenko, Wei Tian

A generalization of the synthetic control method is proposed for a multiple-outcome framework, which improves the reliability of treatment effect estimation. This is done by supplementing the conventional pre-treatment time dimension with the extra dimension of related outcomes in computing the synthetic control weights. The generalization can be particularly useful for studies evaluating the effect of a treatment on multiple outcome variables. To illustrate the method, the effects of non-pharmaceutical interventions (NPIs) are estimated on various outcomes in Sweden in the first3 quarters of 2020. The results suggest that if Sweden had implemented stricter NPIs like the other European countries by March, then there would have been about 70% fewer cumulative COVID-19 infection cases and deaths by July and 20% fewer deaths from all causes in early May, whereas the impacts of the NPIs were relatively mild on the labour market and economic outcomes.

E0764: **How does a better design improve the OLS regression?**

Presenter: **Junho Choi**, Seoul National University, Korea, South

The aim is to demonstrate how balancing makes the inference of the OLS regression robust to model misspecification. First, a general representation of the average treatment effect is derived, which involves the OLS estimand. It decomposes their difference into three components, each bearing on the degree of self-selection, model misspecification, and imbalance in the distribution of control variables between treatment arms. It yields a useful outer bound for the bias of the OLS estimator, whose length is shown to be effectively invariant to misspecification once a better balance is attained. In this sense, better design leads are argued to the robustness of the OLS estimate. Lastly, the findings are extended to the staggered DiD settings.

EO222 Room 203 ADVANCES IN GRAPHICAL MODELS Chair: Federico Camerlenghi

E0527: **Scalable variational Bayes methods for interacting point processes**

Presenter: **Deborah Sulem**, Barcelona School of Economics - Universitat Pompeu Fabra, Spain

Co-authors: Vincent Rivoirard, Judith Rousseau

Multivariate Hawkes processes are temporal point processes extensively applied to model event data with dependence on past occurrences and interaction phenomena, e.g., neuronal spike trains, online messages, and financial transactions. In the nonparametric setting, learning the temporal dependence structure of Hawkes processes is often a computationally expensive task, all the more with Bayesian estimation methods. An efficient algorithm targeting a mean-field variational approximation of the posterior distribution has recently been proposed. Existing variational Bayes inference approaches under a general framework are unified, and it is theoretically analysed under easily verifiable conditions on the prior, the variational class, and the model. Then, in the context of the popular sigmoid Hawkes model, adaptive and sparsity-inducing mean-field variational methods are designed. In particular, a two-step algorithm based on the thresholding heuristic is proposed to select the Granger-causal graph parameter of the Hawkes model. Our approach enjoys several benefits through an extensive set of numerical simulations: it is computationally efficient, can reduce the problem's dimensionality by selecting the graph parameter, and can adapt to the smoothness of the underlying parameter.

E0660: **Bayesian inference of multiple Ising models for heterogeneous data**
Presenter: **Alejandra Avalos Pacheco**, Vienna University of Technology, Austria
Co-authors: Andrea Lazzerini, Monia Lupparelli, Francesco Stingo

Multiple Ising models can be used to model the heterogeneity induced in a set of binary variables by external factors. These factors may influence the joint dependence relationships represented by a set of graphs across different groups. The inference for this class of models is presented, and a Bayesian methodology is proposed based on a Markov Random Field prior to the multiple graph setting. Such prior enables the borrowing of strength across the different groups to encourage common edges when supported by the data. Sparse-inducing priors are employed on the parameters that measure graph similarities to learn which subgroups have a shared graph structure. Two Bayesian approaches are developed for inference and model selection: 1) a Fully Bayesian method for low-dimensional graphs based on conjugate priors specified w.r.t. the exact likelihood, and 2) an Approximate Bayesian method based on a quasi-likelihood approach for high-dimensional graphs where the normalization constant required in the exact method is computationally intractable. The methods' performance is studied and compared with competing approaches through an extensive simulation study. Both inferential strategies are employed to analyze data resulting from two public opinion studies in the US. The first analyzes the confidence in political institutions in different groups divided by the time users spent on web pages. The second studies the opinion on public spending in diverse inter-generational groups.

E0732: **An ANOVA-like decomposition of logistic regression parameters**
Presenter: **Monia Lupparelli**, University of Florence, Italy
Co-authors: Luca La Rocca, Alberto Roverato

A logistic regression setting is considered to study the effect of a focal variable (treatment, risk factor) on a binary outcome given a set of explanatory variables. This effect, measured on the odds-ratio scale, is expected to be heterogenous in multiple models defined by different sets of explanatory variables. Studying this heterogeneity represents a fundamental research question rising in many contexts, i.e., to study distortion of effects, to verify parametric collapsibility with application in precision medicine and mediation analysis. Despite the large interest in this issue, comparing logistic regression coefficients in multiple models, even not nested, is still not a trivial task. Based on a log-hybrid parameterization, an ANOVA-like expansion of the regression coefficient related to the focal variable is provided, where the elements of this expansion are associated with all subsets of the selected variables and are invariant across models. Exploiting this expansion, an exact formula linking the focal regression coefficients in different models has been derived to test hypotheses by answering specific questions simultaneously without fitting multiple logistic models. The resulting class of models has an interesting representation in terms of two-block regression. Results are illustrated for the case of binary variables, but they can also be generalized to the non-binary case.

E1043: **Causal inference for categorical graphical models**
Presenter: **Guido Consonni**, Universita Cattolica del Sacro Cuore, Italy
Co-authors: Federico Castelletti, Marco Della Vedova

A collection of categorical random variables organized in a network is considered. The interest lies in the causal effect on an outcome variable following an intervention on another variable. Conditionally on a Directed Acyclic Graph (DAG), it is assumed that the joint distribution of the random variables can be factorized according to the DAG. The graph is equipped with a causal interpretation through the notion of interventional distribution and the allied do-calculus. The likelihood decomposes into a product of terms, each involving the probability of a node given its parent configurations; the prior, accordingly, is a product of suitably defined Dirichlet distributions. DAG-model uncertainty is taken into account, and a reversible jump MCMC algorithm proposed which targets the joint posterior over DAGs and DAG parameters; from the output, the full posterior distribution of any causal effect of interest is recovered, possibly summarized by a Bayesian Model Averaging (BMA) estimate. The method is validated through simulation studies, wherein the method outperforms alternative state-of-the-art procedures in terms of estimation accuracy. Finally, a dataset on depression and anxiety in undergraduate students is analyzed.

EO143 Room 503 RECENT ADVANCES IN BAYESIAN METHODS: PREDICTION AND CAUSAL INFERENCE Chair: Kenichiro McAlinn

E1234: **Comparing two multivariate stochastic volatility models**
Presenter: **Victor Pena**, Universitat Politecnica de Catalunya, Spain

Multivariate stochastic volatility models are useful for tracking time-varying patterns in covariance structures. Uhlig extended (UE) and beta-Bartlett (BB) processes are especially convenient for analyzing high-dimensional time series because they are conjugate with Wishart likelihoods. It is shown that UE and BB are closely related but not equivalent: their hyperparameters can be matched so that they have the same forward-filtered posteriors and one-step ahead forecasts but different joint (smoothed) posterior distributions. Under this circumstance, Bayes factors cannot discriminate the models, and alternative approaches to model comparison are needed. These issues are illustrated in a retrospective analysis of the volatilities of returns of foreign exchange rates.

E1260: **Inadmissibility and transience**
Presenter: **Kosaku Takanashi**, Riken, Japan
Co-authors: Kenichiro McAlinn

The relation between the statistical question of inadmissibility and the probabilistic question of transience is discussed. The mathematical link between the admissibility of the mean of a Gaussian distribution and the recurrence of a Brownian motion has been proved, which holds for $\mathbb{R}^2$ but not for $\mathbb{R}^3$ in Euclidean space. This result is extended to symmetric, non-Gaussian distributions without assuming the existence of moments. As an application, it is proved that the relation between the inadmissibility of the predictive distribution of a Cauchy distribution with known scale parameter and the transience of the Cauchy process differs from dimensions $\mathbb{R}^1$ to $\mathbb{R}^2$. It is also shown that there exists an extreme model that is inadmissible in $\mathbb{R}^1$.

E1274: **Synthetic control methods through predictive synthesis**
Presenter: **Masahiro Kato**, University of Tokyo, Japan
Co-authors: Akira Fukuda, Kosaku Takanashi, Kenichiro McAlinn, Akari Ohda, Masaaki Imaizumi

Synthetic control (SC) methods have become a vital tool for causal inference in comparative case studies. The main idea of SC methods is to estimate the counterfactual outcomes of a treated unit using a weighted sum of observed outcomes of untreated units. Two novel methods are proposed for SC methods by synthesizing predictive densities. The first method synthesizes predictive densities using Bayesian predictive synthesis (BPS). The proposed method, the Bayesian Predictive SC (BPSC) method, has several advantages over frequentist SC methods. For instance, it can handle issues such as model misspecification and construct confidence intervals. Additionally, covariates can be used as predictors for outcomes by synthesizing them to predict counterfactual outcomes. The second method assumes a mixture model between the densities of treated and untreated units, and SC weights are estimated by density matching. The SC weights can be estimated by matching the higher moments of the treated unit and a weighted sum of untreated units. Using this method, the mean squared error of the counterfactual prediction in experiments is successfully minimized.

E1276: **Bayesian causal synthesis for meta-inference on heterogeneous treatment effects**
Presenter: **Kenichiro McAlinn**, Temple University, United States
Co-authors: Kosaku Takanashi, Shonosuke Sugasawa

A novel Bayesian methodology is proposed to mitigate misspecification and improve estimating treatment effects. A plethora of methods to estimate- particularly the heterogeneous- treatment effect have been proposed with varying success. It is recognized, however, that the underlying data-generating mechanism, or even the model specification, can drastically affect each method's performance without comparing its performance in real-world applications. Using a foundational Bayesian framework, Bayesian causal synthesis is developed, a method that synthesizes several causal estimates to improve overall inference. This process is called meta-inference, as the inference of BCS is occurring above any individual estimate, treating each estimate as data to be updated. A fast posterior computation algorithm is provided, and the proposed method is shown to provide consistent estimates of the heterogeneous treatment effect. Several simulations and an empirical study highlight the efficacy of the proposed approach compared to existing methodologies, providing improved point and density estimation of the heterogeneous treatment effect.

EO313 **Room 603** FINANCIAL MODELLING IN CHANGING MARKET CONDITIONS **Chair: Christina Erlwein-Sayer**

E0735: **Correlation scenarios and correlation stress testing**
Presenter: **Fabian Woebbeking**, IWH, Leibniz Institute for Economic Research, Germany
Co-authors: Natalie Packham

A general approach is developed for stress testing correlations of financial asset portfolios. The correlation matrix of asset returns is specified in a parametric form, where correlations are represented as a function of risk factors, such as country and industry factors. A sparse factor structure linking assets and risk factors is built using Bayesian variable selection methods. The regular calibration yields a joint distribution of economically meaningful stress scenarios of the factors. As such, the method also lends itself to a reverse stress testing framework: using the Mahalanobis distance or Highest Density Regions (HDR) on the joint risk factor distribution allows inferring worst-case correlation scenarios. Examples of stress tests are given on a large portfolio of European and North American stocks.

E0807: **HMM-enhanced LSTM for electricity spot price prediction**
Presenter: **Christina Erlwein-Sayer**, University of Applied Sciences HTW Berlin, Germany
Co-authors: Stefanie Grimm, Tilman Sayer

Electricity spot prices are prone to volatile periods and frequently occurring jumps over time. A prediction of intraday spot prices relies on suitable modelling paradigms to capture these changing dynamics. We develop a long-short-term memory model (LSTM) enhanced by an underlying Hidden Markov model (HMM) to capture regime shifts and ensemble these with the deep learning architecture. Market regimes are adaptively filtered out from the data set and utilised to split the spot price series. We develop an n-state HMM-LSTM which is trained on split regime-specific electricity prices. The prediction is then gained by weighting the LSTM forecast with state probabilities. Combing LSTM with filtered Markov chain probabilities increases the interpretability of predictions. Each activated LSTM is dependent on the filtered state of the underlying market. The proposed model is applied to an extensive data set of German spot prices.

E0842: **Risk factor detection with methods from explainable ML**
Presenter: **Natalie Packham**, Berlin School of Economics and Law, Germany

The importance of risk management in the financial industry has increased rapidly since the financial crisis, particularly regarding financial market stability. A particular focus is stress testing methods, which capture portfolio risk under adverse conditions. Advances in statistical learning and the availability of large, granular data sets offer new methodological possibilities for stress testing. Financial risk management applications such as hedging, scenario analysis and stress testing rely on portfolio models based on risk factors. In addition to observable risk factors, factor models with non-observable, data-based factors offer interesting alternatives. However, the lack of interpretability of the output is limiting. Time-dynamic methods are developed for the interpretability of principal components (PCA) and autoencoders, allowing aggregated risk factors from existing risk factors. This aggregation allows plausibly implementing less granular and even global stress scenarios.

E0417: **Blockchain characteristics and systematic risk: A neural network based factor model for cryptocurrencies**
Presenter: **Alla Petukhina**, HTW Berlin, Germany

A neural network-based factor model is applied to the cryptocurrency market to describe individual asset returns in terms of latent risk factors and time-varying risk exposures. Five contributions are made. First, it is shown that pricing performance is improved by adding nonlinearities to the risk exposures. Second, it is established that the risk dynamics of the cryptocurrency market evolve more quickly than for equity. Third, it is identified that cryptocurrency prices were more predictable before the COVID-19 pandemic than thereafter. Fourth, latent risk factors are found to be related to observables but additionally include idiosyncratic variance. Last, it is observed that asset characteristics that are important for the estimation of risk exposures are commonly found in the literature on observable factor models.

EO082 **Room 604** NEW ADVANCES IN TIME SERIES ECONOMETRICS: THEORY AND APPLICATIONS **Chair: Kun Chen**

E0363: **Negative moment bounds for autocovariance matrices of stationary processes driven by conditional heteroscedastic errors**
Presenter: **Hsueh-Han Huang**, Academia Sinica, Taiwan
Co-authors: Shu-Hui Yu, Ching-Kang Ing

A negative moment bound is established for the sample autocovariance matrix of a stationary process driven by conditional heteroscedastic errors. This moment bound enables us to asymptotically express the mean squared prediction error (MSPE) of the least squares predictor as the sum of three terms related to model complexity, model misspecification, and conditional heteroscedasticity. A direct application of this expression is the development of a model selection criterion that can asymptotically identify the best (in the sense of MSPE) subset AR model in the presence of misspecification and conditional heteroscedasticity. Finally, numerical simulations are conducted to confirm our theoretical results.

E1118: **Does digital finance upgrade trickle-down consumption effect in China?**
Presenter: **Ying Tao**, Sun Yat-sen University, China

The trickle-down consumption and moderating effects of digital finance in China are investigated. Three realistic patterns motivate this study; trickle-down consumption has been confirmed in many developed countries, digital finance facilitates household consumption, and China exhibits income inequality across regions. Using data from the 2017 and 2019 China Household Finance Survey, it is found that Chinese households exhibit trickle-down consumption behaviour and that digital finance development moderates trickle-down consumption; this moderating effect is somewhat heterogeneous by year, geographic location, hukou, and household income. The mechanism analysis results confirm that the relaxation of liquidity constraints is the primary channel through which the positive moderating effect operates. The findings are robust to various tests, and alternative specifications are also discussed. Some policies are suggested to alleviate consumption inequality by developing digital finance and stimulating a natural pattern of trickle-down consumption in China.

E1091: **Sparse causal dynamic linear regression**
Presenter: **Rui Huang**, Nanjing University, China

Longitudinal datasets with multiple time series are common in various fields, such as environmental studies, economics, and finance, motivating the need for effective analytical methods. The dynamic regression model is a powerful tool for exploring linear temporal relationships between variables but may face numerical instability and intractability when handling long time series with multiple leads and lags using the time-domain approach. While the frequency domain approach provides an elegant and efficient alternative solution, the resulting model may suffer from non-causality and non-sparsity, hindering practical utility and interpretability. A novel sparse causal modification is proposed to the general dynamic linear regression model, formulated as a frequency domain functional optimization problem to address these challenges. An accelerated functional proximal gradient descent algorithm is then derived to numerically compute the solution. To further alleviate the computational burden, a computationally efficient one-pass estimation procedure that produces sufficiently good approximate solutions is also proposed when the true underlying model is approximately causal. The theoretical properties of the proposed algorithms are established. The efficacy of the proposed method with both simulation studies and a set of stock index return time series data is showcased.

E1075: **Consistent order selection for ARFIMA processes**
Presenter: **Kun Chen**, Southwestern University of Finance and Economics, China
Co-authors: Ngai Hang Chan, Ching-Kang Ing, Hsueh-Han Huang

Estimating the orders of the autoregressive fractionally integrated moving average (ARFIMA) model has been a long-standing problem in time series analysis. This challenge is tackled by establishing the Bayesian information criterion (BIC) consistency for ARFIMA models with independent errors. Since the memory parameter of the model can be any real number, this consistency result is valid for short memory, long memory and nonstationary time series. Further, the consistency of the BIC is extended to ARFIMA models with conditional heteroscedastic errors, thereby extending its applications to encompass many real-life situations. Finite-sample implications of the theoretical results are illustrated via numerical examples.

EO246 **Room 605** **ESTIMATION, INFERENCE AND FORECASTING IN PANEL DATA ANALYSIS** **Chair: Xiaoyi Han**

E0700: **GMM and root estimations of spatial dynamic panel data models with unknown heteroskedasticity and dominant units**
Presenter: **Chen Yahui**, Xiamen University, China
Co-authors: Han Xiaoyi, Zhang Jiajun

The spatial dynamic panel data model is considered in the presence of dominant units and unknown heteroskedasticity. The dominant units can vary over time, and the number of dominant units can be finite or infinite are allowed. Since the quasi-maximum likelihood estimator (QMLE) is inconsistent under the heteroskedasticity, the generalized method of moments estimator (GMME) and the root estimator (RTE) RE proposed, and the consistency and asymptotic normality of these estimators are established under both scenarios:(1) large n and large T, (2) large n with small T. Our RTE is asymptotically as efficient as the GMME under the heteroskedastic case. Mento Carlo simulations demonstrate that the estimators have satisfactory finite sample performances even if the strength of the dominant units is equal to 1. Finally, an empirical application is presented on the peer effects of firm finance decisions across Chinese listed corporates.

E0713: **Uniform inference for nonparametric panel model with fixed effects**
Presenter: **Nan Liu**, Xiamen University, China

The uniform inference is studied on a structural function $g(.)$ and its functionals in a nonparametric panel data model with individual fixed effects. This nonparametric model relaxes restrictions of time series behaviours by allowing for stationary mixingale or unit root regressors. After removing the individual fixed effects via within-group transformations, a sieve estimation method is proposed, with a sup-norm convergence rate established accordingly. Then, Yurinskii's coupling principle of Gaussian processes and the uniform confidence bands constructed by the sieve score bootstrap method are used to test for linear functionals of $g(.)$. Under the asymptotic framework of an increasing cross-sectional dimension and either a fixed or diverging time dimension, it is proved that the proposed Kolmogorov-Smirnov (sup-type) test has asymptotic uniform size controls and is uniformly consistent. Monte Carlo simulations confirm that the sieve estimator and its uniform confidence bands work well in finite samples. The above uniform inference procedure is also applied in one empirical setting, and some interesting results in nonlinear patterns of elasticity of consumption concerning income shocks are found.

E0784: **Return and volatility forecasting in mixed panels**
Presenter: **Cindy Shin Huei Wang**, HSBC Business School, Peking University, China

A simple pooling prediction is proposed for a mixed panel model including stationary $I(0)$ and $I(d)$ processes via a pool autoregressive approximation (PAR) framework. This PAR-forecasting approach does not require prior information on the exact order and fractional parameter of each series of the mixed panel. It is also shown this approach remains valid when there exist common factors in a mixed panel. Insights from the theoretical analyses are confirmed by a set of Monte Carlo experiments, through which it is demonstrated that the approach outperforms existing forecasting methods. In particular, several controversial arguments in forecasting literature are also justified. Moreover, an empirical application to return and volatility forecasting illustrates the usefulness and feasibility of the forecasting procedure in portfolio settings.

E0981: **Seeding efficient large-scale public health interventions in diverse spatial-social networks**
Presenter: **Xiaoyi Han**, Xiamen University, China
Co-authors: Yilan Xu, Yi Huang, Linlin Fan, Minhong Xu, Song Gao

The selection of target locations for large-scale public health interventions is complex when the take-up of such interventions has peer effects through social networks, and health outcomes have spillover effects through spatial networks. A threshold spatial dynamic panel data (TSDPD) model with endogenized human-virus interactions is developed to address this issue. The model is used to study the initial COVID-19 vaccination rollout in the United States from February 5 to April 15, 2021. Strong peer effects in state vaccination rates arising from the friendship network and strong COVID-19 virus transmission through the travel network are found. Vaccination decreased infections mainly through reduced transmissibility upon passing a full vaccination rate threshold, whereas its impact on the travel network slightly increased infections. Cumulative infections would have been 1.17 million or 24.85% higher if vaccines were not available. Targeting the six most populated US states or most spatially connected is the most clinically effective in reducing cumulative infections by more than 343,000, and targeting the six most socially influential states is the most clinically effective in increasing the national vaccination rate by 3.5 ppts to 18.54%. Targeting the six most socially influential states is more than 20 times as cost-effective as targeting the most populated or most spatially connected ones.

EO199 Room 606 COMPLEX TIME SERIES **Chair: Feiyu Jiang**

E0366: **HAR-Ito models and their high-dimensional statistical inference**
Presenter: **Kexin Lu**, University of Hong Kong, Hong Kong
Co-authors: Huiling Yuan, Yifeng Guo, Guodong Li

Modelling realized volatilities for high-frequency data is an important task in finance and economics. The heterogeneous autoregressive (HAR) model is one of the most popular models used in this area. However, it has three limitations. Firstly, the linear combinations of daily realized volatilities with fixed weights limit its flexibility for different types of assets. Secondly, the high-frequency probabilistic structure of this model, as well as other HAR-type models in the literature, is still unknown. Thirdly, there is no high-dimensional inference tool available for HAR modelling, even though real applications often involve multiple assets. To address these issues, a multilinear low-rank HAR model is proposed using tensor techniques. A data-driven approach is adopted to select the heterogeneous components automatically. HAR-Ito models are also introduced to interpret the corresponding high-frequency dynamics of the proposed model and other HAR-type models. Theoretical properties of high-dimensional HAR modelling are established, and a projected gradient descent algorithm is suggested to search for estimates. The analysis is performed on real data to illustrate the performance of the proposed method.

E0449: **Test of serial dependence or cross dependence for time series with underreporting**
Presenter: **Keyao Wei**, National University of Singapore, Singapore
Co-authors: Yingcun Xia

Testing the serial dependence of a time series or cross dependence of two-time series is essential in time series analysis. Many efficient methods have been proposed for the test when the data is accurately recorded. However, the observed data often systematically deviate from the actual values, a common example being data underreporting in social sciences, ecology, and epidemiology. For these data, it may not be possible to directly make correct inferences using traditional statistical tests. New tests are introduced by using the lag differences of the observed time series, and the statistical consistency of the tests is proven. Further, a block bootstrap method is used to mimic the asymptotic distribution of the test statistics. Numerical experiments show that the proposed tests perform better than existing methods in the case of underreporting. This method has successfully detected important factors for dengue transmission and cardiovascular diseases.

E0483: **Quantiled conditional variance, skewness, and kurtosis by Cornish-Fisher expansion**
Presenter: **Ke Zhu**, University of Hong Kong, Hong Kong

The conditional variance, skewness, and kurtosis play a central role in time series analysis. Some parametric models often study these three conditional moments (CMs) but with two big issues: the risk of model misspecification and the instability of model estimation. To avoid the above two issues, a novel method is proposed to estimate these three CMs by the so-called quantized CMs (QCMs). The QCM method first adopts the idea of Cornish-Fisher expansion to construct a linear regression model based on n different estimated conditional quantiles. Next, it computes the QCMs simply and simultaneously by using this regression model's ordinary least squares estimator without any prior estimation of the conditional mean. Under certain conditions that allow estimated conditional quantiles to be biased, the QCMs are shown to be consistent with the convergence rate $n^{\frac{1}{2}}$. Simulation studies indicate that the QCMs perform well under different scenarios of estimated conditional quantiles. In the application, the study of QCMs for eight major stock indexes demonstrates the effectiveness of financial rescue plans during the COVID-19 pandemic outbreak. It unveils a new "non-zero kink" phenomenon in conditional kurtosis's "news impact curve" function.

E1172: **Testing for innovation symmetry in multivariate generalized autoregressive conditional heteroskedastic models**
Presenter: **Wai-keung Li**, The Education University of Hong Kong, Hong Kong

Testing whether the innovation is symmetric or not is important for multivariate generalized autoregressive conditional heteroskedastic(GARCH) models. The traditional testing methods for univariate models depend on either smoothing or martingale transformation treatment, so their extension to multivariate models is not a straightforward or desirable task. A new consistent test is proposed to examine the symmetry of innovation in multivariate GARCH models using a characteristic measure. Regardless of the dimension of the multivariate GARCH model, the proposed test is easy-to-implement without involving any smoothing treatment. Under certain conditions, the asymptotic null distribution of the test is established. Surprisingly, it is found that the model estimation has a negligible effect on the asymptotic null distribution. Due to this important feature, a simple bootstrap method is provided to compute the critical values of the test. As an extension, similar testing methods for the general multivariate time series models are also applied in the presence of conditional mean.

EO237 Room 701 STATISTICS IN SPORTS **Chair: Marica Manisera**

E0622: **Having a ball: Evaluating scoring streaks and game excitement using in-match trend estimation**
Presenter: **Claus Ekstrom**, University of Copenhagen, Denmark
Co-authors: Andreas Kryger Jensen

Many popular sports involve matches between two teams or players where each team score points throughout the match. While the overall match winner or result is interesting, it conveys little information about the underlying scoring trends throughout the match. Modelling approaches that accommodate a finer granularity of the score difference throughout the match are needed to evaluate in-game strategies and discuss scoring streaks, team strengths, and other aspects of the game. A latent Gaussian process is proposed to model the score difference between two teams and introduce the Trend Direction Index as an easily interpretable probabilistic measure of the current trend in the match as well as a measure of post-game trend evaluation. In addition, the Excitement Trend Index - the expected number of monotonicity changes in the running score difference - is proposed to measure overall game excitement. The proposed methodology is applied to all 1143 matches from the 20192020 National Basketball Association season. It is shown how trends can be interpreted in individual games and how the excitement score can be used to cluster teams according to how excited they are to watch.

E0530: **Evaluating the risk of injury in NBA players**
Presenter: **Ambra Macis**, Universita' degli studi di Brescia, Italy
Co-authors: Marco Sandri, Marica Manisera, Paola Zuccolotto

Injuries commonly occur in sports, and their prevention is important for many reasons. Indeed, injuries have economic implications for teams and may psychologically impact athletes. The survival data analysis framework is considered to evaluate the NBA players' injury risk through the use of statistical models for recurrent events. To this extent, a frailty Cox model has been applied to a unique data set created with a non-trivial harmonization and merging of several data sources: (i) data about all the injuries that occurred from the 2010-2011 season to the 2019-2020 season, (ii) the play-by-play datasets for extracting the information about the amount of minutes played by the players and (iii) an additional dataset with some other specific players' information.

E0353: **Statistically enhanced learning for better predictions**
Presenter: **Christophe Ley**, University of Luxembourg, Luxembourg
Co-authors: Florian Felice, Andreas Groll

Statistically enhanced learning (SEL) is presented, which is a general approach to improve any learning technique, be it statistical or machine learning, by adding highly informative covariates obtained as statistical estimates rather than directly observed. SEL works for any data (tabular, computer vision, text). The general idea is discussed, referring to existing feature extraction methods which actually can be shown to fall under the umbrella of SEL, and its performance is illustrated on both simulated and real data. In particular, it is shown how SEL allows improved predictions of soccer tournaments and discusses how it can be used in sports medicine for injury prevention.

E0664: **Pattern recognition in elite soccer with only a few labeled situations**
Presenter: **Ulf Brefeld**, Leuphana University of Luneburg, Germany

The identification of strategies and tactical patterns is key to pre- and post-match analyses in team sports, and analysts usually spend a great deal of their time watching and annotating video footage. The different ways to automatically annotate patterns of interest so that the analyst can select the most relevant ones for further analyses are discussed. In general, supervised machine learning approaches suggest themselves for this task, but they often require large amounts of labelled situations to successfully learn the target concepts. Since this is often difficult in practice, we will focus on label-efficient approaches, such as self-supervised pre-training, to learn representations of the data, which allows solving pattern and event detection in soccer with only a few annotated situations.

EO306 Room 702 RECENT DEVELOPMENTS IN SPECTRAL IMAGE DATA ANALYSIS Chair: Yunlong Feng

E0232: **Assessment of Raman hyperspectral data from proteins**
Presenter: **Alexander Khmaladze**, SUNY at Albany, United States

Raman spectroscopy offers a non-destructive, label-free approach to classifying biological samples. The Raman effect is a natural phenomenon of inelastic light scattering determined by the vibrational energy levels of specific molecular structures. There are numerous variations on traditional Raman approaches; many require samples to be labelled with a Raman-sensitive compound. However, for monitoring organic samples, unlabeled techniques are ideal. The interpretation of the Raman spectra of biological samples is often dependent on the spectral resolution of the method, with numerous peaks assigned to biological components like DNA, RNA, protein, or lipids based on prior measurements. Often the Raman signature or spectral trends, rather than individual peaks, are utilized to distinguish between biological conditions. It is demonstrated that Raman signature quantitative comparisons based on multivariate analysis and machine learning can be used to detect structural differences in tissues or discriminate healthy tissue regions from disease- or tumour-burdened regions.

E0271: **Boosting climate change adaptation and mitigation by integrated remote sensing image analysis**
Presenter: **Andrea Marinoni**, UiT the Arctic University of Norway, Norway

Remote sensing image analysis enhances our understanding of physical phenomena by combining records acquired from different devices and platforms and allowing higher granularity of information to be extracted about events occurring on the ground. The ability to provide a synoptic view of large areas at regular intervals means remote sensing is fundamental in obtaining a precise characterisation of the nature and extent of dynamic phenomena such as those related to climate change. The proliferation of remote sensing data also gives rise to greater diversity and dimensionality of related datasets. This development allows better monitoring and more precise characterisation of key environmental parameters. Integrating remote sensing images is shown, additional data collected on the ground and via assessment methods, and digital decision-making datasets and products into a coherent multi-scale, multi-level, cross-sectoral climate change adaptation framework that allows the transition towards a climate-neutral and sustainable economy. The result of this approach is the co-creation of regional adaptation pathways with stakeholders, based on the exploration and sequencing of sets of possible actions, to optimise adaptation and mitigation approaches to climate change in a specific region according to the needs of local communities and regional socio-economic development factors.

E0876: **Improved spectral unmixing of highly complex biological fluorescence images using a priori knowledge**
Presenter: **Alex Valm**, State University of New York at Albany, United States

Many biological systems are composed of complex consortia of interacting components. Human dental plaque is known to comprise a community of over 700 different species of bacteria. Knowledge of these communities' spatial structure is critical to understanding disease processes in the oral cavity, including periodontal disease. While it is possible, in principle, to specifically label every species of bacteria in a community with DNA probes, the broad emission spectra of compatible fluorescent dyes prevent the routine use of more than a handful of fluorescent reporters in any single imaging experiment. A constrained unmixing algorithm that incorporates prior knowledge to greatly improve the accuracy in classifying microbial cells according to their taxonomy in images of fluorescently labelled microbial communities is presented. The Sparse and Low-Rank Poisson Regression Unmixing (SL-PRU) approach incorporates multi-penalty terms for rewarding sparseness and spatial correlation of the estimated abundances among spatially correlated pixels. SL-PRU further uses Poisson regression for unmixing instead of least squares regression to better deal with photon shot noise. A general method is proposed to tune the SL-PRU parameter weights and demonstrate improved pixel-wise abundance estimation and endmember classification in complex images of human dental plaque with quantitative morphological analyses.

E0954: **Novel approaches for hyperspectral sensor-based BRDF measurement**
Presenter: **Fadi Kizel**, Technion-Israel Institute of Technology, Israel

Studying the Bidirectional Reflectance Distribution Function (BRDF) of various land cover surfaces is essential in remote sensing. However, measuring the BRDF requires a proper setup and costly special equipment, e.g., the Gonioreflectometer. On the other hand, the recent development in sensing technology allowed for producing spectral cameras that are easy to operate and platforms that combine multiple sensors, for example, a spectroradiometer with an RGB camera. Using these advantages and considering the existing limitation, two novel approaches are presented for measuring the BRDF without requiring unique instruments. The first is suitable for spectral cameras and allows for analyzing the BRDF of surfaces in various scales by acquiring multiple overlapping images in a simple and time-saving way, sampling the desired object's Region Of Interest (ROI) in one image and automatically tracking it in the other images. The second relies on a camera-aided spectroradiometer that simultaneously acquires an RGB image beside the spectral measurement. In both cases, the Structure From Motion (SFM) process is used to retrieve the sensor locations. The results clearly show the highly accurate sensor position derived by SFM, providing zenith angles and distance from the scene's centre with mean errors around one degree and 2.5 centimetres, respectively.

EO067 Room 703 SPATIAL AND NETWORK ECONOMETRICS Chair: Tadao Hoshino

E0584: **Unequally sampled networks: Biases and corrections**
Presenter: **Chih-Sheng Hsieh**, National Taiwan University, Taiwan

Statistical issues arising from networks based on non-representative samples of the population are analyzed. First, the biases in both network statistics and estimates of network effects under unequal probability sampling analytically and numerically are characterized. Sampled network data systematically bias the properties of the population network and suffer from non-classical measurement-error problems when applied as regressors. Apart from the sampling rate and the elicitation procedure, these biases depend in a non-trivial way on which subpopulations are missing with higher probability. A methodology adapting post-stratification weighting approaches is proposed for networked contexts, which enables researchers to recover several network-level statistics and reduce the biases in the estimated network effects. The advantages of the proposed methodology are that it can be applied to network data collected via both designed and non-designed sampling procedures, does not require one to assume any network formation model, and is straightforward to implement. The approach is applied to two widely used network data sets, and it is shown that accounting for the non-representativeness of the sample dramatically changes the results of regression analysis.

E0686: **Sub-model aggregation for scalable spatially varying coefficient modeling**
Presenter: **Daisuke Murakami**, The Institute of Statistical Mathematics, Japan
Co-authors: Shonosuke Sugasawa

The aim is to develop an approach that aggregates/combines global and local sub-models to build a flexible and scalable spatial regression model, including a spatially varying coefficient model. To aggregate sub-models, a generalized product-of-experts method is used, which is widely used in machine learning literature. The aggregated spatial regression model has the following properties: (i) computationally efficient; (ii) each sub-model can be estimated independently; (iii) the marginal likelihood is available in closed form. Owing to (ii), the proposed method is capable of modelling complex spatial patterns. Owing to (iii), its model accuracy is easily compared across specifications. The accuracy and computational efficiency of the proposed method are compared with conventional methods through Monte Carlo simulation experiments. Then the method is then applied to an analysis of residential land prices in Japan.

E0864: **Spatial panel data models with time-varying network structures and multi-dimensional fixed effects**
Presenter: **Zhenlin Yang**, Singapore Management University, Singapore
Co-authors: Xiaoyu Meng

Specification and estimation of spatial panel data models with time-varying network structures and multi-dimensional fixed effects (FEs) are considered. The former capture the (time-varying) endogenous, contextual and correlated social interaction effects, and the latter control the unobserved group, individual, and time heterogeneities. A general strategy is proposed for the identification of common parameters as well as group, individual and time FEs. A simple and general method is introduced for model estimation and inference, where the concentrated quasi-scores (with group, individual and time effects being concentrated out) are adjusted to account for the effect of estimating the incidental FEs. Consistency and asymptotic normality of the proposed estimators are established. Monte Carlo results show excellent finite sample performance of the proposed methods.

E0394: **Causal inference and interpretation of linear social interaction models with endogenous networks**
Presenter: **Tadao Hoshino**, Waseda University, Japan

Causal inference is studied for linear social interaction models in the presence of endogeneity in network formation under a fully heterogeneous treatment effects framework. An experimental setting is considered where individuals are randomly assigned to treatments while no interventions are made on the network structure. It is shown that running a linear regression ignoring the network endogeneity is not problematic for estimating the average direct treatment effect of own treatment, but it leads to both a sample selection bias and a negative weight problem for the estimation of the average spillover effect is proposed to use. To overcome these issues, a potential peer's treatment as an instrumental variable (IV), which is a valid IV for the actual treatment exposure by experimental design. With this IV, several IV-based estimands and demonstrate are examined that they do not suffer from selection bias and have a local average treatment effect-type causal interpretation for the spillover effect.

EO068 Room 705 RECENT ADVANCES IN LARGE-SCALE DATA ANALYSIS Chair: Xiaojun Mao

E0382: **Composite smoothed quantile regression**
Presenter: **Xiaozhou Wang**, East China Normal University, China

Composite quantile regression (CQR) is an efficient method to estimate the parameters of the linear model with non-Gaussian random noise. The non-smoothness of CQR loss prevents many efficient methods from being used. The composite smoothed quantile regression model is proposed, and the inference problem is investigated for a large-scale dataset. The algorithm and theoretical properties are given. Extensive numerical experiments on both simulated and real data are conducted to demonstrate the good performance of the proposed estimator compared to some baselines.

E1310: **Robust personalized federated learning with sparse penalization**
Presenter: **Xiaofei Zhang**, Zhongnan University of Economics and Law, China

Thanks to its advantage in collaborative learning with distributed data, Federated learning (FL) is an emerging topic. Due to the local data-generating mechanism heterogeneity, it is important to consider personalization when developing federated learning methods. A personalized federated learning (PFL) method for addressing the robust regression problem is proposed. Specifically, the aim is to learn the regression weight by solving a Huber loss with the sparse fused penalty. Additionally, the personalized federated learning for robust and sparse regression (PerFL-RSR) algorithm was designed to solve the estimation problem in the federated system efficiently. Theoretically, the convergence property of the proposed PerFL-RSR algorithm is shown, and then the proposed estimator is shown to be statistically consistent. Thorough experiments and real data analysis are conducted to corroborate the theoretical results of the proposed personalized federated learning method.

E1318: **An efficient tensor regression for high-dimensional data**
Presenter: **Yingying Zhang**, East China Normal University, China

Most currently used tensor regression models for high-dimensional data are based on Tucker decomposition, which has good properties but loses its efficiency in compressing tensors very quickly as the order of tensors increases, say greater than four or five. However, for the simplest tensor autoregression in handling time series data, its coefficient tensor already has the order of six. A newly proposed tensor train (TT) decomposition is revised, and then it is applied to tensor regression such that a nice statistical interpretation can be obtained. The new tensor regression can well match the data with hierarchical structures, and it even can lead to a better interpretation of the data with factorial structures, which are supposed to be better fitted by models with Tucker decomposition. More importantly, the new tensor regression can be easily applied to the case with higher-order tensors since TT decomposition can compress the coefficient tensors much more efficiently. The methodology is also extended to tensor autoregression for time series data, and nonasymptotic properties are derived for the ordinary least squares estimations of both tensor regression and autoregression. A new algorithm is introduced to search for estimators, and its theoretical justification is also discussed. The theoretical and computational properties of the proposed methodology are verified by simulation studies, and the advantages over existing methods are illustrated by two real examples.

E1319: **Functional calibration under non-probability survey sampling**
Presenter: **Zhonglei Wang**, Xiamen University, China
Co-authors: Xiaojun Mao, Jae Kwang Kim

Non-probability sampling is prevailing in survey sampling, but ignoring its selection bias leads to erroneous inferences. A unified nonparametric calibration method is offered to estimate the sampling weights for a non-probability sample by calibrating functions of auxiliary variables in a reproducing kernel Hilbert space. The consistency and the limiting distribution of the proposed estimator are established, and the corresponding variance estimator is also investigated. Compared with existing works, the proposed method is more robust since no parametric assumption is made for the selection mechanism of the non-probability sample. Numerical results demonstrate that the proposed method outperforms its competitors, especially when the model is misspecified. The proposed method is applied to analyze the average total cholesterol of Korean citizens based on a non-probability sample from the National Health Insurance Sharing Service and a reference probability sample from the Korea National Health and Nutrition Examination Survey.

EO023 **Room 708** **HIGH-DIMENSIONAL AND SPATIAL FUNCTIONAL DATA** **Chair: Alexander Petersen**

E0269: **Joint estimation of heterogeneous non-Gaussian functional graphical models with fully and partially observed curves**
Presenter: **Eftychia Solea**, Queen Mary University of London, United Kingdom

A new methodology is introduced for estimating undirected graphical models for heterogeneous non-Gaussian multivariate functional data, such as brain activities collected by functional magnetic resonance imaging from a sample of subjects with different subtypes of a neurological disease. The goal of the new model is to estimate robustly a collection of functional graphical models, corresponding to several subpopulations that share some common dependence structure. The model is fitted via a joint estimation method employed with the hierarchical penalty that encourages a common graph structure and individual sparsity. To relax the Gaussian assumption, we consider the functional Gaussian copula graphical model proposed recently and propose the rank-based Kendall's tau correlation operator that extends Kendall's tau correlation coefficient to the functional setting. We establish the concentration inequalities of the estimates and the graph selection consistency for both completely and partially observed data, while allowing the number of functions to diverge to infinity with the sample size. We demonstrate the efficiency of our method through both simulations and an analysis of the fMRI ADHD-200 data set of subjects with inattentive and combined subtypes of ADHD, and control subjects

E0336: **Graphical Gaussian processes for high-dimensional multivariate spatial data**
Presenter: **Abhi Datta**, Johns Hopkins Bloomberg School of Public Health, United States

For multivariate spatial Gaussian process (GP) models, customary specifications of cross-covariance functions do not exploit relational inter-variable graphs to ensure process-level conditional independence among the variables. This is undesirable, especially for highly multivariate settings, where popular cross-covariance functions such as the multivariate Matern suffer from a "curse of dimensionality" as the number of parameters and floating point operations scale up in quadratic and cubic order, respectively, in the number of variables. A class of multivariate "Graphical Gaussian Processes" is proposed using a general construction called "stitching" that crafts cross-covariance functions from graphs and ensures process-level conditional independence among variables. For the Matern family of functions, stitching yields a multivariate GP whose univariate components are Matern GPs and conforms to process-level conditional independence as specified by the graphical model. For highly multivariate settings and decomposable graphical models, stitching offers massive computational gains and parameter dimension reduction. The utility of the graphical Matern GP is demonstrated to jointly model highly multivariate spatial data using simulation examples and an application to air-pollution modelling.

E0958: **Forecasting high-dimensional functional time series: Application to sub-national age-specific mortality**
Presenter: **Han Lin Shang**, Macquarie University, Australia
Co-authors: Ying Sun, Cristian Felipe Jimenez Varon

The focus is on modelling and forecasting high-dimensional functional time series (HDFTS), which can be cross-sectionally correlated and temporally dependent. A novel two-way functional median polish decomposition, which is robust against outliers, is presented to decompose HDFTS into deterministic and time-varying components. A functional time series forecasting method, based on dynamic functional principal component analysis, is implemented to produce forecasts for the time-varying components. By combining the forecasts of the time-varying components with the deterministic components, forecast curves for multiple populations are jointly obtained. Illustrated by the sex- and age-specific mortality rates in the US, France, and Japan, which contain 51 states, 95 departments, and 47 prefectures, respectively, the proposed model delivers more accurate point and interval forecasts in forecasting multi-population mortality than several benchmark methods.

E0971: **Robust functional data analysis for discretely observed data**
Presenter: **Lingxuan Shao**, Fudan University, China
Co-authors: Fang Yao

Robust functional data analysis is considered for discretely observed data with the underlying process having various distributions, such as heavy-tail, skewness, or contaminations. A unified, robust notion of functional mean, covariance, and principal component analysis, are presented, while the existing methods/definitions often differ from each other or concern only fully observed functions (the ideal case). Specifically, the robust functional mean is allowed to be different from its non-robust counterpart and estimated by robust local linear regression, and a new robust functional covariance is defined to share useful properties with the classic version. More importantly, this covariance induces the robust version of Karhunen–Loeve decomposition and corresponding principal components that are useful for dimension reduction. The theoretical results of the robust functional mean, covariance, and eigenfunction estimates, based on pooling discretely observed data (ranging from sparse to dense), are established and coincide with their non-robust counterparts. It is mentioned that the new perturbation bounds for estimated eigenfunctions with indexes allowed to grow with sample size provide a foundation for further modelling based on robust functional principal component analysis.

EO120 Room 709 ADVANCES IN TIME SERIES, RANDOM FORESTS AND CAUSAL INFERENCE Chair: Hiroshi Shiraishi

E0170: **ADCINAR(1) process and bias-correction of some estimators**
Presenter: **Xiaoqiang Zeng**, Hokkaido University, Japan
Co-authors: Yoshihide Kakizawa

The analysis of count time series is an emerging field. During the last four decades, there has been substantial progress on nonnegative integer-valued autoregressive (INAR) type models via the so-called binomial thinning operator (and its variant). The third and fourth auto cumulant (equivalently, central auto moment) functions are derived for the alternative dependent counting nonnegative INAR process of the first-order (ADCINAR(1)). A bias correction using a sample fourth auto moment function is also developed for the commonly used estimators; the Yule-Walker estimator and the conditional least squares estimator. The proposed bias correction, available without computing the closed-form expression for asymptotic expansions of the biases of these two estimators, is practically helpful since the bias formula for the case of the ADCINAR(1) process turns out to be rather complicated.

E0460: **Asymptotic property for generalized random forests**
Presenter: **Hiroshi Shiraishi**, Keio University, Japan
Co-authors: Tomoshige Nakamura, ryuta suzuki

The aim is to develop asymptotic properties of estimators constructed by Generalized Random Forests (GRF), a method to statistically estimate an unknown function defined as a solution to a local estimating equation. By using the theory of empirical processes, the uniform consistency, rate of convergence and weak convergence of the estimator are discussed by GRF.

E0525: **Variable importance measure for generalized random forest**
Presenter: **Tomoshige Nakamura**, Keio University, Japan

The extension of variable importance for random forests to Generalized Random Forests (GRFs) is discussed. GRFs are a method for estimating functional parameters defined as the solution of local estimation equations using random forests. However, unlike the conventional random forest case, the ground truth of the parameters cannot be observed, making it impossible to directly compute the Mean Decrease Accuracy (MDA) from the data. Therefore, an Approximate MDA, which approximates the MDA defined by the functional parameters using a score function, is proposed, and based on this, new Permutation MDA and Noise-up MDA are introduced. As an application, the problem of estimating conditional treatment effects is addressed, and the effectiveness of the proposed methods is demonstrated.

E0573: **Variable importance for random forests: Inconsistency and practical solutions for MDA and Shapley effects**
Presenter: **Clement Benard**, Safran Tech, France

Variable importance measures are the main tools to analyze the black-box mechanisms of random forests. Although the mean decrease accuracy (MDA) is widely accepted as the most efficient variable importance measure for random forests, little is known about its statistical properties. The exact MDA definition varies across the main random forest software. The objective is to analyze the behaviour of the main MDA implementations rigorously. Consequently, their limits are established when the sample size increases. In particular, these limits are broken down into three components: the first two terms are related to Sobol indices, which are well-defined measures of a covariate contribution to the response variance, as opposed to the third term, whose value increases with dependence within covariates. Thus, it is theoretically demonstrated that the MDA does not target the right quantity when covariates are dependent, which has been noticed experimentally. New important measures for random forests are defined to address this issue: the Sobol-MDA and SHAFF. The Sobol-MDA fixes the flaws of the original MDA and is appropriate for variable selection. On the other hand, SHAFF is a fast and accurate estimate of Shapley's effects, even when input variables are dependent. SHAFF is appropriate to rank all variables for interpretation purposes. The consistency of the Sobol-MDA and SHAFF is proved, showing that they empirically outperform their competitors.

EC256 Room 102 FINANCIAL ECONOMETRICS I Chair: Toshiaki Watanabe

E0194: **Robust estimation of the range-based GARCH model: Application for cryptocurrencies**
Presenter: **Piotr Fiszeder**, Nicolaus Copernicus University in Torun, Poland
Co-authors: Marta Malecka

The range-based GARCH model is combined with the modified robust estimation method and suggests a new approach to model the volatility of returns. Thanks to this merger, more information is used, commonly available alongside daily closing prices, i.e., low and high prices. However, the influence of extreme observations is limited in the estimation results. Owing to this, the procedure is not as sensitive to outliers as the maximum likelihood estimation of the range-based models. Introduce the change to the robust method is also proposed, which adds elasticity in treating the outliers and serves to reflect the observations of financial markets, where, after the occurrence of outliers, the volatility persists at an increased level. This method is applied to five selected cryptocurrencies: Bitcoin, Ethereum Classic, Ethereum, Litecoin and Ripple. The forecasts of variance based on the proposed approach are more accurate than forecasts from three benchmarks: the standard GARCH model, the standard range-based GARCH model and the GARCH model with the robust estimation.

E0509: **Intrinsic factor risk premia**
Presenter: **Alberto Quaini**, Erasmus University of Rotterdam, Netherlands
Co-authors: Fabio Trojani, Ming Yuan

Intrinsic factor risk premia are given by the negative factor covariance with the stochastic discount factor projection on the return space. They are well-defined whenever return Sharpe ratios are bounded and equal to zero for any factor uncorrelated with returns. A simple Oracle estimator of intrinsic factor risk premia, which produces reliable inference procedures, is introduced for asset pricing models, including models with factors that are weakly correlated with returns and which consistently selects intrinsically priced factors in finite samples. Using our methodology based on intrinsic risk premia, a broad family of asset pricing models is studied from the factor zoo. In this context, a small set of low-dimensional factor models are detected featuring well-identified factor risk premia and a similarly low degree of misspecification.

E1244: **Modelling volatility with variational inference priciples**
Presenter: **Martin Magris**, Aarhus University, Denmark
Co-authors: Alexandros Iosifidis

Variational Inference (VI) methods are gaining attention and popularity as efficient and practical approaches for performing approximate Bayesian inference in complex models. Concerning various recent VI black-box methods, it is shown that a Bayesian treatment of standard GARCH-like volatility models is immediate and of straightforward implementation. In a study involving 100 stocks and seven volatility models, first, the quality and validity of the VI approximations provided by four optimizers concerning a Monte-Carlo baseline are addressed. Then the impact of the different estimation procedures on both in-sample and out-of-sample performance metrics is discussed. Lastly, it is observed that one-step-ahead volatility forecasts obtained with the above Bayesian methods often outperform their standard likelihood-based counterparts.

E0197: **CAViaR models for value at risk and expected shortfall with long range dependency features**
Presenter: **Gelly Mitrodima**, LSE, United Kingdom
Co-authors: Jaideep Oberoi

Alternative specifications of conditional autoregressive quantile models are considered to estimate value-at-risk (VaR) and expected shortfall (ES). The proposed specifications include a slow-moving component in the quantile process and aggregate returns from heterogeneous horizons as regressors. Using data for ten stock indices over a period that incorporated the COVID-19 pandemic, the performance of the models is evaluated, and the proposed valuable features are found to capture tail dynamics better.

EC262 Room 506 BAYESIAN METHODS **Chair: Boris Choy**

E0992: **Nuisance parameters, modified profile likelihood and Jacobian prior**
Presenter: **Guangjie Li**, Cardiff University, United Kingdom
Co-authors: Roberto Leon-Gonzalez

In a model with nuisance parameters, the maximum likelihood estimators (MLE) of the parameters of interest can be biased. One can reduce the bias due to the presence of the nuisance parameters by removing the $O(1)$ bias of the profile likelihood score. To achieve this, the Jacobian integrated likelihood (JIL) is proposed obtained by using a prior consisting of the Jacobian determinant of the new nuisance parameters, which are functions of the original nuisance parameters and are independent of the dependent variable. The adjusted MPL is proposed, which is easier to be computed and can also remove the $O(1)$ bias of the profile likelihood score. For panel fixed effects models, both the JIL and the adjusted MPL can remove the bias of order $O(T^{-1})$ in the MLE as the cross-sectional size (N) increases. The conditions when the estimators from the adjusted MPL and the JIL are the same and consistent with N being large relative to T are given. Although the adjusted MPL and the JIL do not always exist, one can use their first-order conditions to obtain bias-reduced estimators. The theoretical results are demonstrated by panel binary choice models and dynamic panel linear models with exogenous and predetermined regressors.

E1222: **Langevin-type Monte Carlo algorithms for weakly differentiable non-convex potentials**
Presenter: **Shogo Nakakita**, The University of Tokyo, Japan

Langevin-type Monte Carlo algorithms for distributions with non-convex non-smooth potential functions are considered. If a potential has a weak gradient whose fluctuation within all balls of radius 1 is uniformly bounded, then spherical smoothing can be used to approximate the potential with smoother functions. The spherically smoothed Langevin Monte Carlo algorithm and spherically smoothed stochastic gradient Langevin Monte Carlo one are proposed, and their sampling complexities are discussed.

E1228: **Bayesian estimation of covariate assisted principal regression for brain functional connectivity**
Presenter: **Hyung Gyu Park**, New York University School of Medicine, United States

A Bayesian reformulation of covariate-assisted principal (CAP) regression is presented, which aims to identify components in covariance matrices that are associated with covariates in a regression framework. A geometric formulation and reparameterization of individual covariance matrices in their tangent space are introduced. By mapping the covariance matrices to the tangent space, Euclidean geometry is leveraged to perform posterior inference. This approach enables joint estimation of all parameters and uncertainty quantification within a unified framework, combining dimension reduction for covariance matrices and regression model estimation. To assess the performance of the proposed method, simulation studies are conducted to evaluate its accuracy and efficiency. The method is also applied to analyze associations between covariates and brain functional connectivity, utilizing data from the Human Connectome Project.

E1247: **An efficient Bayesian estimation of nonlinear hierarchical decision models**
Presenter: **Emi Mise**, University of Leicester, United Kingdom
Co-authors: Sanjit Dhami, Ali al-Nowaihi, James Cannam

For estimating decision models such as stochastic cumulative prospect theory (CPT), Bayesian hierarchical modelling is a popular choice. However, standard MCMC algorithms are computationally extremely inefficient or impossible to implement in practice. One, the natural parameters of the statistical model are complex nonlinear functions of the parameters of interest; and two, the dimension of the parameter space grows exponentially with the sample size. An efficient two-stage sampling algorithm is presented, which can cope with any number of parameters. This method is applied to sample the posterior densities of the parameters in two decision theories using experimental data obtained from 556 subjects: CPT and decision by sampling (DbS). The latter has garnered considerable interest but has yet to be tested on experimental data. It is shown that the proposed method works well for both CPT, whose parameters are all continuous and can be transformed to the entire real line, as well as for DbS, which contains a mix of continuous and discrete parameters. It is also demonstrated that DbS has some serious shortcomings despite its early promise.

EC275 Room 704 FORECASTING **Chair: Kaiji Motegi**

E0205: **Alternative percentage error measures for forecasting intermittent and lumpy time series**
Presenter: **Peter Julian Cayton**, University of the Philippines, Philippines

Intermittent and lumpy time series data are kinds of time series in which zero values are frequently observed due to the nature of the observed phenomenon which generated the series. These may be observed from environmental, logistical, and epidemiological processes when the area or scope of the data is small or narrow. Forecast evaluation with intermittent and lumpy time series using percentage errors is complicated as they are difficult to compute given that zero may be a denominator. The research work surveys alternative formulas for percentage error measures in the case of intermittent and lumpy time series and proposes other alternatives for investigation. These measures are assessed using publicly available data on rainfall in key areas within the Philippines, demand logistics data from the R software and other open sources, and COVID-19 reported cases of local government units in the Philippines.

E0415: **An improved test for uniform superior predictive ability**
Presenter: **Verena Monschang**, University of Muenster, Germany
Co-authors: Mark Trede, Wilfling Bernd

The test for uniform superior predictive ability (uSPA) is analyzed, demonstrating that it does not always keep its nominal size. Simulations show that the testing procedure is not able to control the type I error rate. An improved testing approach that keeps the nominal size asymptotically is proposed. Monte Carlo simulations investigate the test's power properties. As an empirical illustration, we compare the predictive ability of two forecasting models.

E0964: **Advancing forecast accuracy analysis: A partial linear instrumental variable and double machine learning approach**
Presenter: **Christoph Schult**, Halle Institute for Economic Research, Germany
Co-authors: Katja Heinisch, Fabio Scaramello

The relationship between forecast accuracy and forecast assumptions is explored using German data and a novel empirical approach. Partial Linear Instrumental Variable (PLIV) regression models are employed, combined with Double Machine Learning (DML) methods, to address high-dimensional nuisance parameters and endogeneity issues. This innovative PLIV-DML framework enables a more complex understanding of the relationships between forecast assumptions and forecasts accuracy than traditional OLS-based analysis. The evaluation sample includes 1460 annual GDP forecasts and various oil, exchange rate, and world trade assumptions. The PLIV-DML model's inherent flexibility allows us to examine two possible violations of assumptions: the rationality of forecasters and the linearity of the data generation process. This research contributes to the field of forecasting by providing a more robust and flexible analysis of forecast accuracy determinants. For instance, the proposed method contributes to the discussion regarding weak instruments and instrumental variables' validity in macroeconomic models. Evidence of a constant underestimation of OLS estimators of the impacts of squared assumption errors of oil price and world trade on squared forecast errors of GDP is found. The insights gained from this study have potential implications for improving economic forecasts' accuracy and understanding underlying forecasting processes.

E1181: **Comparison between forecasting and nowcasting of digital economy**
Presenter: **Pairach Piboonrungroj**, Chiang Mai University, Thailand

Recently, especially since the Coronavirus pandemic (COVID-19), there has been an urgency for real-time (or nearly) economic indicator reports. However, traditional economic indicators such as Gross Domestic Product (GDP) or Employment rate are mostly reported at a low frequency, quarterly or annually. Hence the status quo of economic indicator reporting was found insufficient for economic policy maker to respond to the economic impacts of COVID-19 and the response measure such as lockdowns. Hence, researchers and government agencies attempt to explore alternative indicators to monitor real-time and precise economic situations. Economic trackers proposed using high-frequency data from private sectors and visualizing data on the map by using spatial econometrics. Also, machine learning is employed to analyze alternative data such as Google trends to nowcast the GDP of OECD members. Methods were proposed to nowcast and forecast the digital economy of Thailand using the scope of OECD to track digital economic development. The performance of nowcasting and forecasting are compared and discussed in future research avenues.

EP326 **Room Poster session I** **POSTER SESSION I** **Chair: Cristian Gatu**

E0320: **Unsupervised fuzzy statistical learning and its applications in image segmentation**
Presenter: **Siu Kai Choy**, The Hang Seng University of Hong Kong, Hong Kong
Co-authors: Yee Lam Mo

Fuzzy clustering algorithms, statistical modelling and spatial statistics are popular methodologies in image processing and pattern recognition. However, the literature has not studied the integration of these techniques in image segmentation applications. A robust and effective fuzzy-model-based unsupervised learning algorithm is presented that integrates colour and the wavelet-domain generalized Gaussian density (GGD) statistical model with the fuzzy clustering algorithm combined with neighbouring information for image segmentation applications. Using the GGD statistical model to characterize wavelet subband texture information, the proposed algorithm is particularly effective in segmenting images with complex texture patterns. Comparative experimental results with current existing fuzzy clustering-based approaches show that this methodology achieves remarkable success in image segmentation applications.

E0325: **Optimal investment with return predictability and trading frictions: An asymptotic approach**
Presenter: **Chi Chung Siu**, The Hang Seng University of Hong Kong, Hong Kong
Co-authors: Wing Yan Tsui, Guiyuan Ma

An optimal investment problem of a maximising utility agent with return predictability and trading frictions is studied. Adopting a logarithmic return assumption, the asymptotic expansion around small liquidity costs provides the closed-form expressions for the first-order approximation of the value function and the associated almost-optimal trading strategy. The almost optimal trading strategy indicates that the agent should trade toward the optimal frictionless portfolio instead of directly adopting it. The approximated value function effectively captures the utility loss derived from the agents' inability to adopt the optimal frictionless portfolio over time directly. Finally, the numerical analysis indicates that the agent's utility loss is sensitive to the specifications of the return-predicting factors and that the agent's trading behaviours under the logarithmic return and arithmetic return assumptions can differ remarkably over a medium to a long investment horizon.

E0330: **Subgradient methods for quasi-convex optimization with applications**
Presenter: **Carisa Kwok Wai Yu**, The Hang Seng University of Hong Kong, Hong Kong
Co-authors: Jacky Leung

Subgradient methods form a class of popular and effective iterative algorithms used to solve constrained optimization problems. More recently, they have been developed to solve constrained quasi-convex optimization problems. An implementable subgradient method is proposed where a perturbation of the successive direction is employed at each iteration for solving quasi-convex optimization problems. The proposed subgradient method is applied to solve the Cobb-Douglas production efficiency problem. The numerical study on the efficiency problem shows that the proposed subgradient method outperforms several existing methods, such as the standard, stochastic and primal-dual subgradient methods.

E0406: **Semiparametric survival models with varying coefficients**
Presenter: **Hoi Min Ng**, The Hong Kong Polytechnic University, Hong Kong
Co-authors: Kin Yau Wong

Recent technological advances have made it possible to measure different types of omics features on a large number of subjects. In the studies of chronic diseases such as cancer, it is of great interest to integrate different types of omics features to build a comprehensive understanding of the disease mechanisms. Despite extensive studies on integrative analysis, it remains an ongoing challenge to model the interaction effects among types of omics features due to heterogeneity across data types. A flexible semiparametric varying-coefficient additive hazards model that allows for such interaction effects is developed. The additive hazards model is considered due to its simple interpretation and availability of closed-form estimators. It is noted that most existing kernel-smoothing methods for semiparametric varying-coefficient models are inefficient, as they effectively treat the baseline hazard function as changing with the varying-coefficient variable in the estimation, whereas in the model formulation, the baseline hazard function does not vary. A novel kernel-smoothing method is proposed for estimation which makes use of the fact that the baseline hazard function is shared. The theoretical properties of the proposed estimators are established and their finite-sample performance is investigated by large-scale simulation studies. Applications are provided for a motivating cancer genomic study.

E0572: **Bit-plane probability model and its application in image segmentation**
Presenter: **Kwun Lun Chu**, The Hang Sang University of Hong Kong, Hong Kong

The task of image segmentation is complex in computer vision and has a broad range of applications in fields such as medical imaging and scene analysis. One popular approach often starts with dividing images into non-overlapping blocks and grouping them by similar image features, resulting in an initial segmentation. Boundary correction is then applied to improve this segmentation. Under this framework, selecting the grouping methodology and image features is crucial to the performance of the segmentation result. Traditional approaches in the literature often adopt certain clustering algorithms and rely on the histograms of the image. However, these approaches are very sensitive to initial clusters, and the number of clusters is often unknown, while histograms can be inefficient due to their high dimensionality. To address these challenges and enhance image segmentation accuracy, a novel probability model is proposed to characterize the distributions of image variations based on bit-plane probabilities and dependencies, providing a universal parametric representation that can model any random distribution. In addition, the mathematical optimization framework integrates this model with an agglomerative fuzzy k-means algorithm, incorporating spatial information and a pixel-based boundary localization algorithm for different image segmentation applications. The experimental results demonstrate that the method outperforms current state-of-the-art approaches.

E0617: **Prediction of apartment sale price indices using functional linear models**
Presenter: **Heejin Kim**, Chungnam National University, Korea, South
Co-authors: Eunjee Lee

The aim is to predict apartment sale price indices using linear functional models, a method that analyzes data represented as continuous curves. Specifically, 85 monthly apartment sale price indices from January 2012 to September 2022 provided by the Korea Real Estate Information Center are analyzed. Functional linear models are used to predict the apartment sale price indices, and the accuracy of the predictions is evaluated using root mean square errors (RMSE). We consider the following two methods as competitive models: an autoregressive integrated moving average (ARIMA) and an artificial neural network (ANN) model. The findings suggest that our functional linear model outperforms both the ARIMA and ANN models in terms of overall prediction accuracy. These results have important implications for policymakers and investors seeking to make informed decisions about real estate investments in Korea.

E0747: **Detection of genetic variation related to Alzheimer's disease using functional regression models**
Presenter: **Yoon Seok Lee**, Chungnam National University, Korea, South
Co-authors: Eunjee Lee

Alzheimer's disease (AD) is a neurodegenerative disease that affects the elderly and significantly impacts society. It is the most common form of dementia, accounting for 60 80% of patients, and is a genetic disease influenced by environmental and genetic factors as the nervous system ages. Single nucleotide polymorphisms (SNPs), which are genetic mutations that occur frequently in the human genome, have been identified as a risk factor for the onset of AD. In particular, Apolipoprotein-E is a genetic risk factor for AD, directly related to cognitive decline, and various SNPs such as CR-1 and BIN-1 have been studied as potential risk factors for AD in Genome-Wide Association Studies (GWAS). SNPs are expressed in a densely distributed chromosomal region consisting of as few as 300,000 and as many as 1 million, and adjacent SNPs tend to have similar values. So it's possible to assume that these observed SNPs are discretized observations of a continuous function whose domain is a continuous chromosomal region. Based on this assumption, the SNP data can be considered functional data, a continuum of stochastic sequence data in a continuous space rather than discrete observations. Therefore, a method is proposed to incorporate the SNP data as functional covariates in a logistic regression model with a binary outcome, having Alzheimer's disease or being cognitively normal.

E0848: **Good subsets approach to variable selection**
Presenter: **Neill Smit**, North-West University, South Africa
Co-authors: Riaan de Jongh, Hennie Venter

A recently proposed variable selection procedure is discussed and compared to other well-known variable selection procedures. The lambda-good variable selection procedure is based on selecting good subsets at a specified margin lambda. A subset is said to be good at margin lambda if the associated criterion of fit improves in relative terms if any variable is added, and the criterion of fit deteriorates in relative terms by at least lambda if any variable is dropped. The performance of the lambda-good procedure in terms of variable selection accuracy and computational efficiency will be demonstrated and compared in a simulation study to existing procedures, such as best subsets, forward stepwise selection and the lasso.

Wednesday 02.08.2023 18:00 - 19:15 Parallel Session J – EcoSta2023

EV263 Room 102 BAYESIAN METHODS (VIRTUAL) Chair: Cheng Li

E1068: **Wasserstein convergence in Bayesian deconvolution models**
Presenter: **Catia Scricciolo**, University of Verona, Italy
Co-authors: Judith Rousseau

The purpose is to investigate the multivariate deconvolution problem of recovering the distribution of a signal from i.i.d. observations additively contaminated with random errors having known distribution. We investigate whether a Bayesian nonparametric approach for modelling the latent distribution of the signal can yield inferences with asymptotic frequentist validity under the L^1-Wasserstein metric. For errors with independent coordinates having ordinary smooth densities, we derive an inversion inequality relating the L^1-Wasserstein distance between the distributions of the signal to the L^1-distance between the corresponding mixture densities of the observations. This inequality leads to minimax-optimal rates of contraction for the posterior measure on the distribution of the signal. As an illustration, we consider a Dirichlet process mixture-of-normals prior on the mixing distribution and a Laplace noise. We construct an adaptive approximation of the sampling density by convolving the Laplace density with a well-chosen mixture of normal densities and show that the posterior measure concentrates around the true density at minimax rate, up to a logarithmic factor, in the L^1-metric. The same posterior law automatically adapts to Sobolev regularity of the mixing density, leading to a new Bayesian adaptive estimation procedure for mixing distributions with regular densities, under the L^1-Wasserstein metric.

E1226: **The block-correlated pseudo marginal sampler for state space models**
Presenter: **David Gunawan**, University of Wollongong, Australia
Co-authors: Robert Kohn, Pratiti Chatterjee

Particle Marginal Metropolis-Hastings (PMMH) is a general approach to Bayesian inference when the likelihood is intractable but can be estimated unbiasedly. An efficient PMMH method is developed that scales up better to higher dimensional state vectors than previous approaches. The following innovations achieve the improvement. First, the trimmed mean of the unbiased likelihood estimates of the multiple particle filters is used. Second, a novel block version of PMMH that works with multiple particle filters is proposed. Third, the article develops an efficient auxiliary disturbance particle filter, necessary when the bootstrap disturbance filter is inefficient, but the state transition density cannot be expressed in closed form. Fourth, a novel sorting algorithm, which is as effective as previous approaches but significantly faster than them, is developed to preserve the correlation between the logs of the likelihood estimates at the current and proposed parameter values. The sampler's performance is investigated empirically by applying it to non-linear Dynamic Stochastic General Equilibrium models with relatively high state dimensions and intractable state transition densities and to multivariate stochastic volatility in the mean models.

E0927: **A Bayesian shared-frailty spatial scan statistic model for time-to-event data**
Presenter: **Camille Frevent**, University of Lille, France
Co-authors: Mohamed Salem Ahmed, Sophie Dabo, Michael Genin

Spatial scan statistics are well-known and widely used methods for detecting spatial clusters of events. In the field of spatial analysis of time-to-event data, several models of scan statistics have been proposed. However, these models do not consider the potential intra-unit spatial correlation of individuals nor a potential correlation between spatial units. To overcome this problem, a scan statistic based on a Cox model with shared frailty is proposed that considers the spatial correlation between spatial units. In simulation studies, it has been shown that (i) classical models of spatial scan statistics for time-to-event data fail to maintain the type I error in the presence of intra-spatial unit correlation, and (ii) our model performs well in the presence of both intra-spatial unit correlation and inter-spatial unit correlation. The method has been applied to epidemiological data and the detection of spatial mortality clusters in patients with end-stage renal disease in northern France.

EV278 Room 203 MACHINE LEARNING (VIRTUAL) Chair: Yongdai Kim

E0491: **Learning rates of convolutional neural networks with correntropy induced loss**
Presenter: **Yingqiao Zhang**, Hong Kong Baptist University, Hong Kong

Deep convolutional neural networks are widely used in practice, including image recognition, natural language processing, bioinformatics, and many other fields. Most recent convolutional neural network theory studies are based on the least square loss function. But the least square loss function could not handle the situation well when the noise is heavy-tailed noise which means the noise is only pth moment bounded. Deep convolutional neural networks are investigated with correntropy-induced loss function, assuming the noise is heavy-tailed. It is shown that, with target function in additive ridge functions format, convolutional neural networks followed by one fully connected layer with ReLU activation functions can reach optimal learning rate up to a logarithmic factor, and this rate could circumvent the curse of dimensionality at the same time. In addition, a more general error bound and learning rate are presented when the target function lies in a Sobolev space on the sphere.

E1093: **Machine learning for labor market matching**
Presenter: **Sabrina Muehlbauer**, Institute for Employment Research, Germany
Co-authors: Enzo Weber

A large-scale application is developed to improve the labour market matching process with model- and algorithm-based statistical methods. Matching is defined as a job seeker entering employment by being matched with a specific job. Extensive data is used on employment biographies covering individual and job-related information on employees in Germany. The probability of a job seeker being employed in a certain occupational field is estimated. Thus, a list with any number of job recommendations can be produced for each individual person. The main goal is to improve individual matching by using statistical methods. For this purpose, predictions are made using logit, ordinary least squares regression (OLS), random forest (RF) and k-nearest-neighbours (kNN). The findings suggest that ML performs best regarding the out-of-sample classification error, especially RF. Further, an estimation sample using all spells of persons starting employment performs better than an estimation sample containing only transitions from unemployment into employment. In terms of the unemployment rate, hypothetically, the advantage of ML compared to the common statistical methods could make a difference of 0.3 percentage points.

E1133: **A hybrid two-step approach for assessing the probability of training needs on artificial intelligence systems**
Presenter: **Sabrina Maggio**, University of Salento, Italy
Co-authors: Veronica Distefano, Sandra De Iaco

Artificial Intelligence (AI) represents the core of many technologies and in the last few years, it has become more and more crucial in helping and enhancing decision-making processes. A wide variety of research studies has been developed in AI, covering many different areas, from Health to Agriculture, from Industry to Information Technology. Nevertheless, only a few works have focused on the impact of applying AI on people's confidence and their reflections on training needs. The novelty of this study concerns the introduction of a hybrid two-step approach based on machine learning and multilevel modeling to assess the effect of people's awareness, attitude and trust in AI on the probability of training needs. In particular, the Boruta Random Forest algorithm will be applied to identify the key determinants of training needs in AI in eight European countries to be included in the multilevel logit model. Then, the probability of European citizens' educational needs in AI will be computed and analyzed with respect to gender.

EO173 **Room 04** SURVIVAL ANALYSIS: THEORY AND METHODS **Chair: Takeshi Emura**

E0636: **Hierarchical penalized distributional regression models for survival data**
Presenter: **Kevin Burke**, University of Limerick, Ireland
Co-authors: Fatima-Zahra Jaouimaa, Il Do Ha

A distributional regression approach is taken to analyse survival data, whereby explanatory variables can enter the hazard regression model through its scale and shape parameters; this enables flexible modelling beyond proportional hazards. A penalised hierarchical likelihood estimation approach is adopted to facilitate automatic variable selection and account for hierarchical data structures (e.g. clustered clinical trials). This very general procedure applies an adaptive lasso to both the scale and shape hazard parameters while incorporating correlated bivariate frailty in both parameters. The estimation and inferential performance of the proposal are investigated using simulation studies; the method on a real data example is demonstrated.

E0681: **An approach for long-term survival data with dependent censoring**
Presenter: **Silvana Schneider**, Federal University of Rio Grande do Sul, Brazil

In long-term studies, some causes of censoring are generally falsely assumed to be independent, leading to bias being neglected. Therefore, a likelihood-based approach is proposed for long-term clustered survival data, which is suitable to accommodate the dependent censoring. The association between lifetimes and dependent censoring is accommodated through the conditional approach of the frailty models. The marginal distributions can be adjusted assuming Weibull or piecewise exponential distributions, respectively. A Monte Carlo Expectation-Maximization algorithm is developed to estimate the proposed estimators. The simulation study results show a small relative bias and coverage probability near the nominal level, indicating that the proposed approach works well. Moreover, the model identifiability is assured once data has a cluster structure. Finally, the survival times of free-ranging dogs from West Bengal, India, collected between 2010 to 2015, are analyzed, and it is concluded that survival time (death due to natural cause) is negatively correlated to dependent censoring (missing cause).

E0846: **Analysis of doubly truncated data**
Presenter: **Carla Moreira**, University of Minho, Portugal

Truncation is a well-known phenomenon that may be present in observational studies of time-to-event data. For example, when the sample restricts to those individuals with event falling between two particular dates, they are subject to selection bias due to the simultaneous presence of left and right truncation, also known as interval sampling, leading to a double truncation. When time-to-event data is doubly truncated, the sampling information includes the variable of interest X and left-truncation and right-truncation variables U and V, but the observable population reduces to those individuals for which the variable of interest lies between left-truncation and right-truncation variables. In this case, both large and small values of X are observed in principle with a relatively small probability. The problem of estimating the distribution of X and other related curves such as kernel density and kernel hazard functions, using nonparametric and semiparametric approaches, from a set of iid triplets with distribution of (X, U, V) given the double truncation restriction will be presented. Several epidemiological scenarios where the effect of ignoring double truncation appears in practice will be reported. Possible limitations of the nonparametric and semiparametric estimators will be discussed.

EO027 **Room Virtual R01** ASSET PRICING, AND RISK ATTITUDES TOWARDS RARE DISASTERS **Chair: Go Charles-Cadogan**

E1256: **On the predictability of stock returns using predictive equity analytics with dynamic state space**
Presenter: **Mark Zanecki**, IHA Consultants, United States

Predictive equity analytics with a dynamic state space framework is introduced to measure both the static and dynamic components of equity return processes. A state is associated with each close return followed by state compaction to enable matrix eigenvalue quantification. The framework allows for identifying and quantifying outlier processes separate from the core return process. Short-run predictability is demonstrated where the signal is sufficiently intense above the background, and long-run predictability uses maximal eigenvalue as well as the first occurrence of meaningful change in eigenvalue. Predictive equity analytics with dynamic state space amplifies spectral analysis by providing a dynamic Kalman-like filter as a first step.

E1320: **Robo-advising under rare disasters**
Presenter: **Jiawen Liang**, University of Glasgow, United Kingdom
Co-authors: Cathy Yi-Hsuan Chen, Bowei Chen

Robo-advisors provide automated portfolio management services to investors, and their growth has been unprecedented in the past few years. However, empirical evidence shows that robo-advisors underperformed during the recent COVID-19 pandemic. This may be because rare disasters are highly unlikely to occur and yet have a huge impact on financial markets. A novel computational framework is developed to improve the performance and robustness of robo-advising in the presence of rare disasters. It integrates reinforcement learning with importance sampling. Instead of sampling transition probability from a ground-truth probability distribution, it is sampled from a proposal distribution, where the event of interest occurs more frequently. The proposed algorithm is validated by data covering the 2008 financial crisis and the COVID-19 pandemic, showing superior performance over benchmarked methods. The algorithm is model-free and reduces the variance of value estimates through importance sampling. In addition to methodological contributions, the study contributes to the growing literature on robo-advising by considering rare events.

E1255: **Quantitative easing of fear during rare disasters**
Presenter: **Go Charles-Cadogan**, University of Leicester, United Kingdom

A natural experiment is conducted in which the US Federal Reserve money supply M1SL (M1SL) response to the Great Recession of 2008 is a control, and its M1SL response to the exogenous COVID-19 pandemic event is a treatment for the Great Lockdown of 2020. Both recession periods are matched on market crashes from rare disasters, almost identical VIX scores, similar jumps in risk aversion, similar U-shape patterns in US Treasury yield curves, and similar intertemporal marginal rate of substitution (IMRS) behaviour. The main difference is the Feds' unprecedented M1SL treatment of the Great Lockdown compared to the Great Recession of 2008 (and other rare disasters). A novel time-varying stochastic discount factor (SDF) is introduced, which admits explosions based on fear of catastrophic loss in financial markets, and it disentangles risk aversion from fear of loss. It is found that even though both rare disasters were matched on risk attitudes and consumption behaviour, the unprecedented increase in M1SL treatment during the Great Lockdown is aliasing for fear of loss like that observed during the Great Recession. The treatment is tantamount to latent risk substitution that attenuated the SDF during the V-shaped Great Lockdown recession. The risk substitution is confirmed by a difference-in-difference analysis.

EO144 Room Virtual R02 ENVIRONMENTAL DATA MODELING, PREDICTION AND RISK ASSESSMENT Chair: Stefano Rizzelli

E0856: **Bayesian inference and probabilistic forecasting for the peaks over threshold approach**
Presenter: **Simone Padoan**, Bocconi University, Italy
Co-authors: Stefano Rizzelli, Clement Dombry

The focus is on the Peaks Over Threshold (POT) method, arguably the most popular approach in the univariate extreme values literature. Many useful inferential procedures for estimating extreme events have been developed in the last decades. To the best of our knowledge, the more ambitious and challenging problem of proper probabilistic forecasting of future extremes has received little or no attention to date. A prior distribution that allows handling the issues arising from using the Generalised Pareto distribution as a misspecified model for the inference regarding the POT method is discussed, and the asymptotic theory of the posterior distribution follows from it is investigated. The primary purpose of risk analysis is the prediction of future extreme events. The problem of probabilistic forecasting of future extremes is addressed by adopting the Bayesian paradigm. Starting from our proposed Bayesian procedure, the posterior predictive distribution of a future unobserved excess above a high threshold is specified, and it is shown that it is Wasserstein consistent concerning the true distribution of such an unobserved future observation.

E0837: **Predicting risks of temperature extremes using large-scale circulation patterns with r-Pareto processes**
Presenter: **Jonathan Koh**, University of Bern, Switzerland

Many severe weather patterns in the mid-latitudes have been found to be connected to a particular atmospheric pattern known as blocking. This pattern obstructs the prevailing westerly large-scale atmospheric flow, changing flow anomalies in the vicinity of the blocking system to sustain weather conditions in the immediate region of its occurrence. Blocking presence and characteristics are thus important for the development of temperature extremes, which are rarely isolated in space, so one must not just account for their occurrence probabilities and intensities but also their spatial dependencies when assessing their associated risk. A methodology is proposed that does so by combining tools from the spatial extremes and machine learning to incorporate 500hPa geopotential (Z500) anomalies over the North Atlantic and European region as covariates to predict surface temperature extremes. This involves fitting Generalized r-Pareto processes with appropriate risk functionals to daily high-impact positive and negative temperature anomaly events across central Europe from 1979-2020, using loss functions motivated by extreme-value theory in a gradient boosting algorithm. It is checked by simulation that the model parameters are identifiable and can be estimated adequately. It is found which circulation patterns in the Euro-Atlantic sector are most important in determining the characteristics of these extremes and showing how they affect them.

E0579: **Wasserstein distributional data analysis with application to wind forecasting**
Presenter: **Matteo Pegoraro**, Aalborg University, Denmark

Many environmental data come in the form of probability distributions. Due to the uncertainties involved in environmental processes, data are often aggregated to compare different phenomena better. Also, predictions are better understood regarding probability distributions over the possible outcomes. It is thus very important to develop techniques which can be used to solve data analysis problems related to distributional data sets. A framework is presented for principal component analysis and regression when statistical units are probability measures considered with the Wasserstein metric.

E0585: **Bayesian multi-species N-mixture models for large scale spatial data in community ecology**
Presenter: **Michele Peruzzi**, Duke University, United States

Community ecologists seek to model the local abundance of multiple animal species while considering that observed counts only represent a portion of the underlying population size. Analogously, modelling spatial correlations in species' latent abundances is essential when attempting to explain how species compete for scarce resources. A Bayesian multi-species N-mixture model with spatial latent effects to address both issues is developed. On the one hand, the model accounts for imperfect detection by modelling local abundance via a Poisson log-linear model. Conditional on the local abundance, the observed counts have a binomial distribution. On the other hand, a directed acyclic graph restricts spatial dependence is let to speed up computations and recently developed gradient-based Markov-chain Monte Carlo methods are used to sample a posteriori in the multivariate non-Gaussian data scenarios in which it is interested. The model is illustrated on synthetic data and data from the North American Breeding Bird Survey.

EO035 Room 201 ADVANCES IN MODELLING ORDINAL AND MIXED-TYPE DATA (VIRTUAL) Chair: Monia Ranalli

E0765: **Estimation and accuracy evaluation of cyber-risk prioritization for threat intelligence**
Presenter: **Mario Angelelli**, University of Salento, Italy
Co-authors: Serena Arima, Christian Catalano, Enrico Ciavolino

The pervasive diffusion of interconnected Information and Communication Technologies (ICTs) is driving the need to properly assess cyber risk and prioritize counteractions, aiming to better manage resources for cybersecurity in the prevention of cyber incidents. The increasing complexity of digital systems requires flexible models for extracting the information necessary to ensure data integrity, confidentiality, and availability. Motivated by this need, a new statistical methodology is introduced to support cyber-vulnerability prioritization expressed in terms of ordinal data, which assesses the severity of a vulnerability. The new method combines a non-parametric regression model and a new accuracy index meeting operative requirements often encountered in the cybersecurity field. Specifically, the methodology uses mid-quantile regression as a robust approach for ordinal severity assessments, and the proposed accuracy measure enjoys invariance properties for consistent ranking derivation. The proposed model is tested on simulated and real data, which are obtained from the fusion of different databases providing relevant information on exploiting cyber vulnerabilities. The proposal is compared with alternative methods (ordered logit, linear regression for rank-transformed variables) to evaluate the potential advantages of this approach, the domains where these advantages are significant, and their interpretation for threat intelligence.

E1037: **An analysis of mine-related insurance data using a compositional approach**
Presenter: **Francesco Porro**, Universita degli Studi di Genova, Italy

One of the most critical issues faced by insurance companies is the continuous monitoring of the number of accidents experienced by their insured clients. An analysis of variables that can provide information on the forecasting of such events is therefore very worthy for insurance companies, especially if it is performed through innovative approaches and novel methodologies. A dataset provided by the US Mine Safety and Health Administration regarding the characteristics of a set of US mines, including the number of occurred accidents in the time range 2013-2016, is considered. Since most of the considered variables are either compositional or categorical, the analysis should be executed by means of the appropriate techniques. In particular, compositional data can be investigated by using the correct statistical tools in order to have reliable results. Following this approach, a compositional (CoDa) analysis is performed, taking into account that the relevant information conveyed by the compositional data is in the proportions among the parts and not in their absolute values or in their sum.

E1124: **Parsimonious and semi-constrained models for clustering mixed-type data through a composite likelihood approach**
Presenter: **Monia Ranalli**, Sapienza University of Rome, Italy
Co-authors: Roberto Rocci

Twelve parsimonious models for clustering mixed-type (ordinal and continuous) data are proposed. Ordinal and continuous data are assumed to follow a multivariate finite mixture of Gaussians. Two main closely related issues should be faced with when the dimensionality of the data increases: the number of parameters increases exponentially; a large number of ordinal variables makes the full maximum likelihood estimation infeasible. To solve the first issue, the model should be more parsimonious in terms of the number of parameters to estimate. At this aim, a general class of eight parsimonious mixture models for mixed-type data are defined by imposing a factor decomposition on component-specific covariance matrices. The loadings and variances of error terms of the factor model may be constrained to be equal or unequal across mixture components. To add some extra flexibility to maintain a certain degree of parsimony, four further models are defined, where the latent factors in each cluster are the same but with different variances. A nice feature of these semi-constrained models is that, under mild conditions, the factors are unique. In other terms, it is impossible to rotate the factors as in the classical factor analysis model. To solve the second issue, a composite likelihood approach is adopted. Estimates computation is carried out using an EM-type algorithm based on composite likelihood. The proposal is evaluated through a simulation study and an application to real data.

EO021 Room 506 ML IN ELECTRICITY PRICING, ACTUARIAL LOSS RESERVING AND EFFICIENCY ANALYSIS Chair: Yuning Zhang

E1041: **New Zealand electricity price forecasting: An analysis of statistical and machine learning models with feature selection**
Presenter: **Gaurav Kapoor**, Auckland University of Technology, New Zealand
Co-authors: Nuttanan Wichitaksorn

An empirical comparison is presented for statistical and machine learning models for daily electricity price forecasting in the New Zealand electricity market. The effectiveness of GARCH and SV models and their t-distribution variants when paired with feature selection techniques, including LASSO, mutual information, and recursive feature elimination, are demonstrated. A key aspect of the study is the inclusion of a diverse set of explanatory variables in all models. These models are compared against a range of popular machine learning models, including LSTM, GRU, XGBoost, LEAR, and a four-layer DNN, where the latter two are considered benchmarks. The results reveal that GARCH and SV models, particularly their *t* variants, perform exceptionally well when paired with feature selection techniques and explanatory variables. In most scenarios considered, these models outperform machine learning models when coupled with LASSO feature selection. This contribution provides a comprehensive evaluation of the performance of different models and feature selection techniques for electricity price forecasting in the New Zealand electricity market. The best-performing model improves the symmetric mean absolute percentage error (sMAPE) and means absolute scaled error (MASE) by 2% to 3% over the LEAR benchmark model, highlighting the practical relevance of the findings.

E0648: **Stochastic loss reserving with long short term memory**
Presenter: **Yuning Zhang**, The University of Sydney Business School, Australia
Co-authors: Boris Choy, Junbin Gao

A flexible mixture density network (MDN) approach for stochastic loss reserving in general insurance is proposed. To model the temporal sequences of claim losses presented in the run-off triangle, a special bi-directional Long Short Term Memory (2DLSTM) is employed. Unlike the original bi-directional Recurrent neural network (biRNN) and bi-directional LSTM, which train the model in both forward and backward time directions, the 2DLSTM uses input information from the top and left neighbours of the run-off triangle during the training procedure. This allows the proposed approach to capture both the accident period and development period dynamics in loss reserving.

E1033: **Bayesian nonparametric machine learning approach for efficiency analysis**
Presenter: **Zheng Wei**, Texas A&M University, United States
Co-authors: Huiyan Sang, Nene Coulibaly

The stochastic frontier model is widely used in economics, finance, and management to estimate the production function and efficiency of a firm or industry. In the current literature on stochastic frontier analysis, parametric forms of the production function, such as Cobb-Douglas and translog, are often assumed a priori without validation, which may suffer from model misspecification and lead to biased efficiency estimates. To address this issue, a new stochastic frontier model built upon a monotone-constrained nonparametric production function is proposed via an extension of the monotone Bayesian Additive Regression Tree (MBART) framework, which allows for greater flexibility in modelling the production function with uncertainty measure while accounting for complex relationships between high dimensional data and variable selection. The performance of the proposed model was illustrated through simulation studies and a real data application.

EO066 Room 604 RECENT ADVANCES IN ECONOMETRICS Chair: Seok Young Hong

E0746: **Augment large covariance matrix estimation with auxiliary network information**
Presenter: **Shuyi Ge**, University of Nankai, China

The aim is to incorporate auxiliary information about the structure of significant correlations into the estimation of static high-dimensional covariance matrices. With the development of machine learning techniques such as textual analysis, granular linkage information among firms that used to be notoriously hard to get is now becoming available to researchers. The proposed method provides an avenue for combining those auxiliary network information with traditional statistical regularization models, mainly banding and thresholding, to improve the estimation of a large covariance matrix. Simulation results show that the proposed adaptive correlation thresholding and banding methods generally perform better in the estimation of covariance matrices than competitors, especially when the true covariance matrix is sparse and the auxiliary network contains genuine information. Empirically, the method is applied to the estimation of the covariance matrix of asset returns to attain the global minimum variance portfolio.

E0565: **A dynamic semiparametric characteristics-based model for optimal portfolio selection**
Presenter: **Shaoran Li**, Peking University, China
Co-authors: Oliver Linton, Chaohua Dong, Gregory Connor

A two-step semiparametric methodology is developed for portfolio weight selection for characteristics-based factor-tilt and factor-timing investment strategies. It is built upon the expected utility maximization framework. Asset returns are assumed to obey a characteristics-based factor model with time-varying factor risk premia. It is proved under our return-generating assumptions that an approximately optimal portfolio can be established using a two-step procedure in a market with a large number of assets. The first step finds optimal factor-mimicking sub-portfolios using a quadratic objective function over linear combinations of characteristics-based factor loadings. The second step dynamically combines these factor-mimicking sub-portfolios based on a time-varying signal, using the investors' expected utility as the objective function. A two-stage semiparametric estimator is developed and implemented. It is applied to CRSP (Center for Research in Security Prices) and FRED (Federal Reserve Economic Data) data, and excellent in-sample and out-sample performance consistent with investors' risk aversion levels is found.

E1097: **Nonparametric range-based estimation of integrated variance with episodic extreme return persistence**
Presenter: **Shifan Yu**, Lancaster University, United Kingdom
Co-authors: Yifan Li, Ingmar Nolte, Sandra Nolte

A new nonparametric estimator of integrated variance is developed based on intraday candlestick information (high, low, open, and close prices in short time intervals). This range-return difference volatility (RRDV) estimator is robust to short-lived extreme return persistence hardly attributable to the diffusion component, such as gradual jumps and flash crashes. By modelling such sharp but continuous price movements following two recent influential works, it is shown that RRDV can provide consistent estimates with relatively small variances. Simulation results demonstrate the reliability of the proposed estimator in practice with some finite-sample refinements. An empirical illustration of volatility forecasting shows the RRDV-based Heterogeneous Autoregressive (HAR) model performs well relative to existing procedures according to standard out-of-sample loss functions.

EO045 Room 605 HAWKES PROCESSES IN ECONOMETRICS AND STATISTICS Chair: Yoann Potiron

E0182: **Estimation of integrated intensity in Hawkes processes with time-varying baseline**
Presenter: **Olivier Scaillet**, University of Geneva and Swiss Finance Institute, Switzerland
Co-authors: Yoann Potiron, Seunghyeon Yu

Transaction times are modelled as a Hawkes process with a time-varying baseline and a general kernel. The baseline is assumed to be the sum of a deterministic seasonal component and a stochastic Itô semimartingale with possible jumps. In *mixed* asymptotics, a nonparametric estimation of the integrated intensity is provided. In addition, the integrated intensity is decomposed as a sum of the contributions of the seasonal and random parts.

E0413: **Modelling financial volatility with quadratic Hawkes**
Presenter: **Cecilia Aubrun**, Ecole Polytechnique, France
Co-authors: Jean-Philippe Bouchaud, Michael Benzaquen

Hawkes processes have been used in various fields, from seismology to finance to model endogenous dynamics. Those stochastic processes are particularly well suited to the problem because the feedback effect is explicitly described via a kernel that weights the influence of past events on the frequency of occurrence of future events. Non-linear extensions of Hawkes processes allow one to combine both excitatory and inhibitory effects and can describe an even broader range of phenomena: brain function, financial markets activity and volatility, and seismologic activity. Financial markets offer a prolific playground to study non-linear Hawkes. One special class of such non-linear processes, called quadratic Hawkes, was introduced and studied to model price movements. On top of the standard Hawkes feedback, a signed process (price changes in this context) also contributes to the current activity rate in a quadratic way. QHawkes is particularly interesting because it allows us to reproduce the stylized facts of financial time series: clustering of activity, fat-tailed distribution of financial returns, and time asymmetry. When considering several assets, one observes additional stylized facts: cross feedback effects between several financial products (cross leverage and Zumbach effects) and simultaneous price jumps of several assets, co-jumps. Considering those, we generalize the QHawkes by extending it in multi-dimensions.

E0652: **Mutually exciting point processes with latency**
Presenter: **Yoann Potiron**, Keio University, Japan
Co-authors: Vladimir Volkov

A novel statistical approach to estimating latency, defined as the time it takes to learn about an event and generate a response to this event, is proposed. Our approach only requires a multidimensional point process describing the arrival time of events, which circumvents the use of more detailed datasets which may not even be available. We consider the class of parametric Hawkes models capturing clustering effects in which latency is defined as a known function of kernel parameters, typically the mode of kernel distribution. Relying on a realistic mixture of generalized gamma kernels, the estimation of model parameters is performed via quasi-maximum likelihood and the feasible limit theory with in-fill asymptotics is derived. As a byproduct, asymptotic theory for a latency estimator, defined as the function of parameter estimates and two tests, is deduced. Numerical studies corroborate the theory. Latency estimates for the US and Canadian stock exchanges vary between 2 and 13 milliseconds from 2020 to 2021. The US firms are found to be more involved in relative latency competition, implying different risk appetites for firms with different latencies.

EO152 Room 702 RANDOM MATRIX THEORY FOR COMPLEX DATA (VIRTUAL) Chair: Jesus Arroyo

E0314: **Resurrecting pseudoinverse: Asymptotic properties of large Moore-Penrose inverse with applications**
Presenter: **Nestor Parolya**, Delft University of Technology, Netherlands
Co-authors: Taras Bodnar

High-dimensional asymptotic properties of the Moore-Penrose inverse of the sample covariance matrix are derived, i.e., when the number of variables p is larger than the sample size n. The convergence results related to the traces of weighted moments of the Moore-Penrose inverse matrix, which involve both its eigenvalues and eigenvectors, are proved. Previous findings are extended in several directions: (i) first, the population covariance matrix is not assumed to be a multiple of the identity; (ii) second, the assumptions of normality are not used in the derivation, only the existence of the 4th moments is required; (iii) third, the asymptotic properties of the weighted moments are derived under the high-dimensional asymptotic regime, when both p and n approaches infinity such that p/n tends to a constant c>1. The findings allow the construction of the optimal linear shrinkage estimators for large precision matrix, beta-vector in the high-dimensional linear models and minimum L2 portfolio. Finally, the finite sample properties of the derived theoretical results are investigated via an extensive simulation study.

E1099: **Spiked eigenvalues of high-dimensional sample autocovariance matrices: CLT and applications**
Presenter: **Yanrong Yang**, The Australian National University, Australia
Co-authors: Han Xiao Han

High-dimensional autocovariance matrices play an important role in dimension reduction for high-dimensional time series. The central limit theorem (CLT) is established for spiked eigenvalues of high-dimensional sample autocovariance matrices developed under general conditions. The spiked eigenvalues are allowed to go to infinity in a flexible way without restrictions in divergence order. Moreover, the number of spiked eigenvalues and the time lag of the autocovariance matrix under this study could be either fixed or tending to infinity when the dimension p and the time length T go to infinity together. As a further statistical application, a novel autocovariance test is proposed to detect the equivalence of spiked eigenvalues for two high-dimensional time series. Various simulation studies are illustrated to justify the theoretical findings. Furthermore, a hierarchical clustering approach based on the autocovariance test is constructed and applied to clustering mortality data from multiple countries.

E1189: **Covariance and autocovariance estimation on a Liouville quantum gravity sphere in a functional context**
Presenter: **Andrej Srakar**, University of Ljubljana, Slovenia

Research on spherical random fields and their applications has become an important part of probability, statistics and mathematical physics. Approaches are extended to study anisotropic spherical random fields previously unaddressed in this context area in random geometry, namely

Liouville quantum gravity (LQG) spheres. The quantum Liouville theory was introduced in 1981 as a model for quantizing the bosonic string in the conformal gauge and gravity in two space-time dimensions. Liouville measure is formally the exponential of the Gaussian free field (GFF), and it is possible to study in depth its properties about SLE curves or geometrical objects in the plane that can be constructed out of the GFF. The problem of estimation of Green-type covariance is studied, and autocovariance functions of a continuous Gaussian free field are defined on an LQG sphere. Their estimators are proposed within a functional data analysis context and study their asymptotics, including their computational aspects. In an application, data on sea surface temperature anomalies (temperature and salinity of the upper 2000 m of the ocean) is studied and recorded by Argo floats. In conclusion, extensions are discussed in many areas of studying spherical random fields and their relationship to probability and random geometry in a functional context.

EO192 Room 703 ADVANCES IN TIME SERIES AND SPATIAL DATA ANALYSIS Chair: Soudeep Deb

E0656: **Semiparametric estimation method for quantile coherence with an application to financial time series clustering**
Presenter: **Cristian Felipe Jimenez Varon**, King Abdullah University of Science and Technology, Saudi Arabia
Co-authors: Ying Sun, Ta-Hsin Li

In multivariate time series analysis, the coherence measures the linear dependency between two-time series at different frequencies. However, real data applications often exhibit nonlinear dependence in the frequency domain. Conventional coherence analysis fails to capture such dependency. Quantile coherence, conversely, characterizes nonlinear dependency by defining the coherence at a set of quantile levels. Although quantile coherence is a more powerful tool, its estimation remains challenging due to the high level of noise. A new semi-parametric estimation technique is proposed for quantile coherence. The method uses the parametric form of the spectrum of the vector autoregressive (VAR) model as an approximation to the quantile spectral matrix, along with a nonparametric smoother. For each quantile level, the VAR parameters from the quantile periodograms are obtained, and then, using the Durbin-Levinson algorithm, the initial estimate of quantile coherence is calculated. Finally, it is smoothed across quantiles with a nonparametric smoother. Numerical results show outperformance over nonparametric methods. It is shown that quantile coherence-based time series clustering has advantages over ordinary coherence. For applications, the identified clusters of financial stocks by quantile coherence with a market benchmark are shown to have an intriguing and more accurate structure of diversified investment portfolios that may be used by investors to make better decisions.

E0688: **Bayesian spatial modeling for data fusion adjusting for preferential sampling**
Presenter: **Paula Moraga**, King Abdullah University of Science and Technology (KAUST), Saudi Arabia

Spatially misaligned data are becoming increasingly common due to data collection and management advances. A Bayesian geostatistical model for combining data obtained at different spatial resolutions is presented. The flexible model can be applied in preferential sampling and spatiotemporal settings. The model assumes that underlying all observations, a spatially continuous variable can be modelled using a Gaussian random field process. The fast inference is performed via the integrated nested Laplace approximation (INLA) and the stochastic partial differential equation (SPDE) approaches. A new SPDE projection matrix for mapping the Gaussian Markov random field from the observations to the triangulation nodes is proposed to allow spatial data fusion. The performance of the new approach by means of simulation and air pollution applications is shown. The approach presented provides a useful tool in a wide range of situations where information at different spatial scales needs to be combined.

E0821: **Recent advances in parameter estimation and model selection of differential equations with application for digital twins**
Presenter: **Itai Dattner**, University of Haifa, Israel

A Digital Twin is a virtual representation of a physical entity designed to capture its dynamics in as much detail as possible. This allows for better design, prediction, and control of physical entities over their lifetime. Differential equations are a powerful mathematical tool to describe dynamic processes and thus play an essential role in the development of Digital Twins. Recent advances will be reviewed in parameter estimation and model selection of differential equations, discussing statistical theory, methodology, and applications to real data.

EO220 Room 705 ESTIMATION FOR MODELS WITH COMPLEX STRUCTURAL DATA Chair: Qian Lin

E1027: **Semiparametric efficient estimation of genetic relatedness with double machine learning**
Presenter: **Niwen Zhou**, Beijing Normal University, China

Double machine learning procedures are proposed to estimate genetic relatedness between two traits in a model-free framework. Most existing methods require specifying certain parametric models involving the traits and genetic variants. However, the bias due to model misspecification may yield misleading statistical results. Moreover, the semiparametric efficient bounds for estimators of genetic relatedness are still lacking. Semiparametric efficient and model-free estimators are developed, and valid confidence intervals are constructed for two important measures of genetic relatedness: genetic covariance and genetic correlation, allowing both continuous and discrete responses. Based on the derived efficient influence functions of genetic relatedness, a consistent estimator of the genetic covariance is proposed as long as one of the genetic values is consistently estimated. The data of two traits may be collected from the same group or different groups of individuals. Various numerical studies are performed to illustrate the introduced procedures. Also proposed procedures are applied to analyze Carworth Farms White mice genome-wide association study data.

E1261: **Quantile regression with asynchronous longitudinal data**
Presenter: **Xuerui Li**, Beijing Normal University, China
Co-authors: Yanyan Liu, Yuanshan Wu, Lixing Zhu

In many biomedical longitudinal studies, time-dependent responses and covariates are observed asynchronously within subjects. And the biomedical data often presents heteroscedasticity with outliers and a skewed distribution in response. Due to the fact that quantile regression is generally robust in handling skewed responses in heteroscedastic data and flexible to accommodate covariate-response relationships, we consider quantile regression modelling to include time-invariant and time-varying coefficients for such longitudinal data. Asymptotic properties are established, including consistency and weak convergence. Simulations studies suggest the good finite-sample performance of the proposed method. The practical example concludes more comprehensive results under the proposed estimation when comparing directly using synchronous data analysis methods and mean regression.

E1272: **Sliced inverse regression with large structural dimension**
Presenter: **Qian Lin**, Tisnghua University, China
Co-authors: Dongming Huang, Songtao Tian

The central space of a joint distribution (X,Y) is the minimal subspace $\mathcal{S}$ such that $Y \perp\!\!\!\perp X \mid P_{\mathcal{S}}X$ where $P_{\mathcal{S}}$ is the projection onto $\mathcal{S}$. Sliced inverse regression (SIR), one of the most popular methods for estimating the central space, often performs poorly when the structural dimension $d = dim(\mathcal{S})$ is large (e.g., ≥ 5). It is demonstrated that the generalized signal-noise-ratio (gSNR) tends to be extremely small for a general multiple-index model when d is large. Then the minimax rate for estimating the central space is determineovered a large class of high dimensional distributions with large structural dimension d (i.e., there is no constant upper bound on d)in the low gSNR regime. This result extends the existing minimax rate results for estimating the central space of distributions with fixed d to that with large d and clarifies that the decay of signal strength causes the degradation in SIR performance. The technical tools developed here might be of independent interest for studying other central space estimation methods.

EC325 Room 02 NON-PARAMETRIC HYPOTHESIS TESTING **Chair: Michele Guindani**

E0323: **On the (in)admissibility of generalized permutation tests**
Presenter: **Nick Koning**, Erasmus University Rotterdam, Netherlands

Recently, novel size-controlling permutation tests for exchangeability have been introduced. Unlike the traditional tests that require a uniform distribution of permutations, these new tests can be based on any distribution of permutations. While this opens up a wealth of possibilities to construct new tests, the power properties of these tests are yet to be understood. The main contribution is to demonstrate that this generalization is inadmissible by showing that a size-controlling traditional permutation test with a randomized level dominates any generalized permutation test. In addition, conditions under which the traditional permutation test dominates its randomized-level counterparts are provided. Finally, these tests and results from permutations to arbitrary compact groups and weaker null hypotheses incorporating the chosen test statistic are extended.

E0340: **Testing second-order stochastic dominance**
Presenter: **Tommaso Lando**, University of Bergamo, Italy
Co-authors: Sirio Legramanti

Second-order stochastic dominance (SSD) is probably the main ordering relation in fields such as economics and finance. A nonparametric test is proposed that measures deviations from the null assumption of SSD. Critical values of the test may be obtained using bootstrap procedures. Asymptotic properties are studied from a theoretical perspective and by means of simulation studies.

E0454: **Stochastic monotonicity of statistical functionals with testing application**
Presenter: **Paulo Oliveira**, University of Coimbra, Portugal
Co-authors: Idir Arab, Tommaso Lando

A general family of stochastic orders encompassing quite a few common ordering notions is introduced. This notion is used to derive stochastic monotonicity relations for adequate families of statistical functionals, including several popular measures such as generalized entropy or Gini indices. Moreover, this approach also reveals to be useful for deriving finite sample properties of nonparametric goodness-of-fit tests.

EC328 Room 03 MISSING DATA **Chair: Masayuki Hirukawa**

E0184: **Regression estimation for continuous time functional data processes with missing at random response**
Presenter: **Mohamed Chaouch**, Qatar University, Qatar
Co-authors: Naamane Laib

Nonparametric kernel of a generalized regression function based on an incomplete sample $(X_t, Y_t, \zeta_t)_{t\in[0,T]}$ copies of a continuous-time stationary and ergodic process (X, Y, ζ) are estimated. The predictor X is valued in some infinite-dimensional space, whereas the real-valued process Y is observed when the Bernoulli process $\zeta = 1$ and missing whenever $\zeta = 0$. Uniform almost sure consistency rate and the evaluation of the conditional bias and asymptotic mean square error are established. The asymptotic distribution of the estimator is provided with a discussion on its use in building asymptotic confidence intervals. To illustrate the performance of the proposed estimator, a first simulation is performed to compare the efficiency of discrete-time and continuous-time estimators. A second simulation is conducted to discuss the selection of the optimal sampling mesh in the continuous-time case. Then, a third simulation is considered to build asymptotic confidence intervals. Finally, an application to financial time series is used to study the performance of the proposed estimator in terms of point and interval prediction of the IBM asset price log returns.

E1156: **Missing endogenous variables in conditional moment restriction models**
Presenter: **Antonio Cosma**, University of Bergamo, Italy
Co-authors: Andrei Kostyrka, Gautam Tripathi

The focus is on estimating finite dimensional parameters identified via a system of conditional moment equalities when at least one of the endogenous variables (which can either be endogenous outcomes, endogenous explanatory variables, or both) is missing for some individuals in the sample. The semiparametric efficiency bound is derived for estimating the parameters, and it is used to demonstrate that if all of the endogenous variables in the model are missing, then estimation using only the validation subsample (the subsample of observations for which the endogenous variables are nonmissing) is asymptotically efficient. An estimator based on the full sample is also proposed that achieves the semiparametric efficiency bound. A simulation study reveals that the estimator can work well in medium-sized samples and that the resulting efficiency gains (measured as the ratio of the variance of an efficient estimator based on the validation sample and the variance of the estimator) are comparable with the maximum gain the simulation design can deliver.

E1160: **Approaches for handling missing values and their impacts on biological inferences: A molecular rate case study**
Presenter: **Zeny Feng**, University of Guelph, Canada
Co-authors: Jacqueline May, Sarah Adamovicz

The association between the variation of molecular evolutionary rates and species traits is a prevalent pattern across the Tree of Life. However, analyses that aim to identify such trait-rate associations are often limited in scope due to the missing values in trait data. A common practice of using a complete-case analysis by removing species with missing values will reduce the sample size and analysis power. In the study of the correlates between the molecular rates of cytochrome c oxidase subunit I (COI) and traits of ray-finned fishes, using the complete-case data, the sample size is reduced to 20% of the original dataset. Missing data imputation offers an alternative that helps to retain sample size, but its accuracy is subject to the choice of imputation methods. The impact of imputation on biological inferences remains largely unexplored, with much focus on imputation accuracy using simulated datasets. Here, we propose a real data-based simulation strategy to select the best-suited method to impute the missing values in the fish trait data. Phylogeny information of multiple nuclear genes will also be used for imputing the missing trait values. Among datasets resulting from different missing data handling approaches, their resulting distributions are compared for each trait. The trait-rate association analysis will also be performed using these datasets. Results will be compared to assess their impacts on the significance level of the trait-rate association.

EC304 Room 503 ASSET ALLOCATION **Chair: Yifeng Guo**

E0277: **High dimensional portfolio selection with cardinality constraints**
Presenter: **Yifeng Guo**, The University of Hong kong, China

The expanding number of assets offers more opportunities for investors but poses new challenges for modern portfolio management (PM). As a central plank of PM, portfolio selection by expected utility maximization (EUM) faces uncontrollable estimation and optimization errors in ultrahigh-dimensional scenarios. Past strategies for high-dimensional PM mainly concern only large-cap companies and select many stocks, making PM impractical. A sample-average-approximation-based portfolio strategy is proposed to tackle the difficulties above with cardinality constraints. Our strategy bypasses estimating the mean and covariance of the Chinese walls in high-dimensional scenarios. Empirical results on the S&P 500 and Russell 2000 shows that an appropriate number of carefully chosen assets leads to better out-of-sample mean-variance efficiency. On Russell 2000, our best portfolio profits as much as the equally-weighted portfolio but reduces the maximum drawdown and the average number of assets by 10% and 90%, respectively. The flexibility and the stability of incorporating factor signals for augmenting out-of-sample performances are also demonstrated. Our strategy balances the trade-off among the return, the risk, and the number of assets with cardinality constraints. Therefore, a theoretically sound and computationally efficient strategy is provided to make PM practical in the growing global financial market.

E1211: **Dynamic dependence of equity market factors and empirical analysis of tail risk parity factor portfolio**
Presenter: **Kakeru Ito**, Nissay Asset Management, Japan
Co-authors: Naoki Makimoto

Since the global financial crisis 2008, factor investing has become popular among investment managers. However, not much is known about dependence across factors. Furthermore, several papers have recently pointed out that factor returns tend to deteriorate simultaneously, leading to worse portfolio performance than expected. The dynamic dependence of Japanese equity market factors with dynamic skewed t copula is estimated, and the results with those estimated by traditional multivariate copulas are compared. The impact of capturing the dependence across the factors on the performance of the tail risk parity factor portfolio is also considered. Four important results are found. First, capturing the dynamic linear correlations enhances the performance of the risk parity factor portfolio compared with the static risk parity factor portfolio. Secondly, the tail risk parity factor portfolio tends to outperform the risk parity factor portfolio, especially Sortino ratio is improved. Third, considering the asymmetric tail dependence improves the maximum drawdown of the tail risk parity factor portfolio. Fourth, introducing the CVaR targeting strategy also improves the maximum drawdown of the tail risk parity factor portfolio. These findings indicate that managing the downside risk of the tail risk parity factor portfolio with dynamic skewed t copula is useful.

E1087: **Intertemporal substitution and risk with multiple assets and quantile preferences**
Presenter: **Hirofumi Ota**, Rutgers University, United States
Co-authors: Antonio Galvao, Luciano De Castro, Daniel Nunes

Novel economic models and econometric methods are developed to jointly identify and estimate intertemporal preference parameters and risk attitudes. First, an intertemporal consumption model with multiple assets is suggested using dynamic quantile preferences that incorporate the elasticity of intertemporal substitution, risk attitude, and discount factor. Interesting explicit expressions are obtained for the value function, the optimal asset allocation and consumption, as well as for the consumption path. Next, the quantile Euler equation is derived, where it is shown that when at least two returns are available, one is able to separately identify the risk attitude, which is measured by the quantile tau, and the elasticity of intertemporal substitution. Finally, based on the identification result from the quantile Euler equation, a new econometric theory is developed for estimating the parameters of interest. The proposed new semiparametric two-step estimator is based on sample splitting and a smoothed l1-norm and allows for estimation of the risk attitude specified by quantile tau and elasticity of intertemporal substitution. A non-standard cubic root asymptotic theory is established, where the limiting distribution is the maximizer of the convolution of two-scaled Brownian motion with quadratic drift. Finally, an inference procedure is suggested via subsampling, providing simulation results to investigate the finite-sample performance.

EC280 Room 603 RISK ANALYSIS **Chair: Malika Hamadi**

E1151: **Exploring and forecasting stochastic risk of payments series for loan and pension scheme**
Presenter: **Nurul Hasanah Uswati Dewi**, Universitas Hayam Wuruk Perbanas, Indonesia
Co-authors: Khreshna Syuhada

A series of payments may be intended to repay a loan or to deposit a pension scheme. It is usually assumed that such payments are non-stochastic, i.e., an individual is making payments regularly with a certain interest rate and no overdue payments or even failing to make payments. However, this assumption may not always hold in actual situations, as there may be unforeseen circumstances that could impact one's ability to make payments on time. A model is constructed for a series of stochastic payments. In particular, the potential risk(s) of failing is explored to repay a loan or make a deposit in a pension scheme. The corresponding statistical distribution is developed, and forecasting future risk is carried out. This can help individuals and institutions to manage the risk, improving their financial stability and ensuring long-term sustainability. Numerical analysis of both simulation and real data is also used to illustrate the model and future risk forecast.

E1205: **Credit risk in microcredit markets**
Presenter: **Malika Hamadi**, Birmingham Business School, United Kingdom
Co-authors: Andreas Heinen, Jeremie Juste

The recent microfinance repayment crises and the growing importance of the microfinance sector as a share of GDP in many developing countries raise concerns about its stability. The determinants of systemic credit risk in microcredit markets in 37 countries from 2000 to 2014 are studied by investigating the dependence of the portfolio-at-risk (PAR) of microfinance institutions (MFIs) within a country. To that end, a panel model of equidependent Gaussian copulas is introduced, where the microfinance sector in a given country year is viewed as a portfolio of MFIs, which becomes riskier when the dependence among their PAR increases. Thus, more dependence on nonperforming loans of MFIs in a country is an indicator of systemic risk in the sector. This methodology allows us to control for country-fixed effects to correct the incidental parameter bias that the non-linear model is subject to. It is found that measures of competition, the sector's fast growth, commercialization, and the average interest rate charged by MFIs increase risk, while the proportion of women borrowers in the sector and country-level remittances reduce it. Further, the probability is computed that a proportion of at least 20% of the MFIs in a country is experiencing serious repayment problems. It is shown that this probability increased prior to the outbreak of a repayment crisis in several countries.

E1149: **Stock liquidity and value at risk for options**
Presenter: **Shih-Ping Feng**, Feng Chia University, Taiwan

Traditional Value at Risk (VaR) estimates for options assume that the underlying stock has perfect liquidity, but in reality, investors trade stock with liquidity risk. Empirical studies have clearly documented that the liquidity risk of the underlying assets plays a role in the distribution of stock and option returns. A method is presented for calculating liquidity-adjusted VaR estimates for options that account for the imperfect liquidity of the underlying asset. The imperfect liquidity of an asset is assumed to result from an imbalance in market demand and supply. Empirically, the predictive accuracy of the proposed liquidity-adjusted VaR estimates for options is compared with traditional VaR estimates. The empirical results show that the proposed liquidity-adjusted VaR estimates achieve a relatively good fit, especially when the underlying stock exhibits lower levels of liquidity.

EC315 Room 606 ECONOMETRIC THEORY Chair: Teppei Ogihara

E1104: **Ratio tests using the Cauchy distribution: A simple principle**
Presenter: **Uwe Hassler**, Goethe University Frankfurt, Germany
Co-authors: Mehdi Hosseinkouchack

A testing principle is introduced by building on two weighted partial sample sums. Under general assumptions, both sums are asymptotically normal. Upon normalization and orthogonalization, the ratio thus converges to the standard Cauchy distribution. Critical values and p-values are hence readily available, and local power can be computed against specific alternative hypotheses. At the same time, a potential nuisance scaling parameter cancels from the ratio making these Cauchy tests self-normalizing. wo examples are discussed: a test for the null of zero mean random variables and a test for the null of a unit root in time series. Both examples result in a limiting Wiener process. Asymptotic local power is evaluated against different alternatives. The weights in the numerator and denominator are from the Karhunen-Loeve expansion of the Wiener process. The power crucially hinges on the specific weighting schemes used to compute the test statistics. Finally, this Cauchy test principle is carried to a multivariate framework of several correlated samples. Here, the cross-covariances between the samples reduce to one scaling parameter that cancels from the Cauchy ratios due to self-normalization. The tests are robust with respect to cross-dependence without the need to estimate nuisance parameters.

E1022: **Uniform convergence rates for nonparametric estimators of a density function when the density has a known pole**
Presenter: **Sorawoot Srisuma**, National University of Singapore, Singapore

The aim is to study the uniform convergence rates of nonparametric estimators for a probability density function and its derivatives when the density has a known pole. Such a situation arises in some structural micro econometric models, e.g., in auction, labour, and consumer search, where uniform convergence rates of density functions are important for nonparametric and semiparametric estimation. Existing uniform convergence results based on Rosenblatt's kernel estimator are derived under the assumption that the density is bounded. They are not applicable when there is a pole in the density. The pole nonparametrically is treated, and various kernel-based estimators are shown can attain any convergence rate that is slower than the optimal rate (when the density is bounded) uniformly over an appropriately expanding support under mild conditions.

E1210: **Hypothesis testing for mediation effects in a generalized regression model**
Presenter: **Jung Hyub Lee**, The University of Texas at Austin, United States

A unifying framework is considered for testing causal mediation effects in nonlinear models. A generalized linear-index model is introduced and extended to incorporate endogenous treatments and endogenous mediators. This model does not impose parametric assumptions on the error terms. A kernel-weighted Kendall's tau is leveraged to test the significance of the indirect effect of endogenous treatments on the outcome variable of interest mediated by endogenous mediators. The proposed semiparametric model allows for treatments and mediators to be discrete, continuous, or neither of these two (e.g., censored or truncated). Two distinct kernel-weighted Kendall's tau statistics will be constructed that capture the effect of the treatment on the mediator, and the mediator on the outcome variable of interest, respectively. However, it turns out that typical joint hypothesis tests using these statistics demonstrate the severely low size of a test and low test power. A similar problem also has been reported when using standard linear causal mediation models. To tackle the problem, a test method is leveraged that is a 'nearly similar powerful' that gives a size of the test sufficiently close to the desired level. For empirical illustration, we consider the British Household Panel Survey data to assess the effect of education level on social functioning mediated by annual individual income.

EC299 Room 701 STOCHASTIC VOLATILITY Chair: Toshiaki Watanabe

E0256: **Idiosyncratic volatility factor and macroeconomic risks**
Presenter: **Yangming Bao**, Capital University of Economics and Business, China
Co-authors: Ying Lun Cheung

An econometric framework is proposed to extract the common factor among assets' idiosyncratic volatilities documented in the literature lately. The idiosyncratic volatility factor (IVF) is shown to have a superb asset pricing ability and command a negative price of risk as predicted by theory. Exploring the link between IVF and macroeconomic risks, a strong positive relationship is found between IVF and macroeconomic uncertainty. Using a fractionally cointegrated VAR model, further the long-run equilibrium relation is uncovered between IVF and the market volatility and the impact of disequilibrium on macroeconomic uncertainty.

E0374: **Infinite sparse factor stochastic volatility model**
Presenter: **Martina Zaharieva**, CUNEF SL, Spain

A sparse factor multivariate stochastic volatility model is proposed, in which the sparsity of the loading matrix is achieved by introducing the Indian buffet process, a Bayesian nonparametric, prior to defining a distribution over infinite binary matrices. The benefit of the infinite-dimensional latent process is twofold. First, inducing sparsity prior reduces the dimensionality of the problem, and second, the number of active factors is determined by the data itself and a priori set to infinity. Both the diagonal elements of the covariance matrix of the idiosyncratic term and the active factors follow univariate stochastic volatility processes. Each latent volatility is sampled independently and in parallel by means of a particle filtering and smoothing technique based on a simulated likelihood. The model is applied to a cross-section of five international stock market indices.

E1218: **Stochastic volatility model with range-based correction and leverage**
Presenter: **Yuta Kurose**, University of Tsukuba, Japan

Contemporaneous modeling of asset return and price range within the framework of stochastic volatility with leverage are presented. A new representation of the probability density function for the price range is provided, and its accurate sampling algorithm is developed. A Bayesian estimation using the Markov chain Monte Carlo (MCMC) method is provided for the model parameters and unobserved variables. MCMC samples can be generated rigorously, despite the estimation procedure requiring sampling from a density function with the sum of an infinite series. The empirical results obtained using data from the U.S. market indices are consistent with the stylized facts in the financial market, such as the existence of the leverage effect. In addition, to explore the model's predictive ability, a model comparison based on the volatility forecast performance is conducted.

EC190 Room 704 TIME SERIES II Chair: Zudi Lu

E0260: A new SSA-based procedure for detecting structural changes in a time series

Presenter: **Adelaide Freitas**, University of Aveiro, Portugal
Co-authors: Alberto Silva

Some procedures adopted to detect eventual structural changes in a time series using Singular Spectral Analysis consist of applying a single decomposition method to two different trajectory matrices (base and test) iteratively throughout the series. Then, distances between some eigenvectors and an appropriate subspace are computed and compared in each iteration. A method is proposed to assess differences when two decomposition methods (robust and ordinary) are applied to the same trajectory matrix. These differences will be more accentuated when there is an eventual change in the direction of some principal components (eigenvectors) in case of interrupting the linear recurrent formula. One advantage of this strategy lies in the possibility of interpretation in terms of the principal components that the visualization of the results provides.

E0868: Bias-reducing penalization for the Whittle likelihood

Presenter: **Francesca Papagni**, Free University of Bolzano, Italy
Co-authors: Davide Ferrari, Greta Goracci

Whittle likelihood estimation is a widely used and computationally efficient approach to approximate the Gaussian likelihood in the frequency domain. However, it produces biased parameter estimates when the samples are relatively small. The empirical adjustment is considered for the Whittle likelihood function, which is shown to reduce the size of the asymptotic bias of the resulting estimates. The validity of the approach is shown through asymptotic calculations and Monte Carlo experiments. The examples focus on long-memory models, for which the Whittle likelihood represents one of the most commonly used estimation methods. The numerical findings show that significant bias reductions in small samples characterize the adjusted Whittle estimates. As an illustration, the new methodology is applied to the analysis of the Southern Oscillation Index data based on a relatively short series, which is relevant for predicting year-to-year climate variation in the global climate leading to floods, droughts and other natural disasters.

E0261: A two-stage maximum entropy approach for time series regression

Presenter: **Pedro Macedo**, University of Aveiro, Portugal
Co-authors: Jorge Duarte, Maria Costa, Mara Madaleno

The maximum entropy bootstrap for time series is a powerful technique that creates a large number of replicates, as elements of an ensemble, for inference analysis. As an alternative to the use of some traditional estimators, generalized maximum entropy is proposed for the estimation of parameters in all the models generated by the maximum entropy bootstrap. A simulation study suggests that generalized maximum entropy is competitive (in a mean squared error loss sense) with some traditional estimators when the models are reasonably well-conditioned and is superior in ill-conditioned scenarios. Empirical applications on energy markets and climate change science are provided to illustrate the procedures where maximum entropy is used both in data replication (maximum entropy bootstrap) as well as in parameter estimation (generalized maximum entropy).

EC268 Room 708 COMPLEX DATA ANALYSIS Chair: Ray-Bing Chen

E0396: Finite mixture model based on the GSMMGN family with several interval censoring

Presenter: **Ruijie Guan**, Beijing University of Technology, China
Co-authors: Tsung-I Lin, Weihu Cheng

The generalized scale mixtures of mixture generalized normal (GSMMGN) distribution is presented, which is a versatile family of distributions capable of modeling data with diverse and flexible shapes. A novel finite mixture model based on the GSMMGN class of distributions with several interval censoring (FM-GSMMGN-SIC) is established, which provides a basic framework for modeling complex data exhibiting multimodality, large skewness, heavy tails, leptokurtic or platykurtic behaviors, and missing values simultaneously. A variant of the EM-type algorithm is formulated by combining the reparameterization technique and profile likelihood approach (PLA) with the classical Expectation Conditional Maximization (ECM) algorithm for parameter estimation of the proposed model. This approach with analytical expressions in the E-step and tractable M-step can greatly enhance the computational speed and efficiency of the algorithm. Some simulation studies are conducted to assess the performance of the proposed algorithm, and the results show satisfactory outcomes for several artificial datasets. Moreover, the feasibility and practical usefulness of the proposed methodology are illustrated through real data analysis.

E0701: Nested Grassmannians for dimensionality reduction with applications

Presenter: **Chun-Hao Yang**, National Taiwan University, Taiwan

Recently, the nested structure of Riemannian manifolds has been studied in the context of dimensionality reduction as an alternative to the popular principal geodesic analysis (PGA) technique, for example, the principal nested spheres. A novel framework is proposed for constructing a nested sequence of homogeneous Riemannian manifolds. Common examples of homogeneous Riemannian manifolds include the spheres, the Stiefel manifolds, and the Grassmann manifolds. In particular, it is focused on applying the proposed framework to the Grassmann manifolds, giving rise to the nested Grassmannians (NG). An important application in which Grassmann manifolds are encountered is planar shape analysis. Specifically, each planar (2D) shape can be represented as a point in the complex projective space, a complex Grassmann manifold. Some salient features of the framework are: (i) it explicitly exploits the geometry of the homogeneous Riemannian manifolds, and (ii) the nested lower-dimensional submanifolds need not be geodesic. With the proposed NG structure, algorithms are developed for the supervised and unsupervised dimensionality reduction problems respectively. The proposed algorithms are compared with PGA via simulation studies and real data experiments and are shown to achieve a higher ratio of expressed variance compared to PGA.

E1066: Tukey's depth for object data

Presenter: **Sara Lopez Pintado**, Northeastern University, United States
Co-authors: Xiongtao Dai

A novel exploratory tool for non-Euclidean object data is developed based on data depth, extending celebrated Tukey's depth for Euclidean data. The proposed metric halfspace depth, applicable to data objects in general metric spaces, assigns to data points depth values that characterize the centrality of these points concerning the distribution and provides an interpretable centre-outward ranking. Desirable theoretical properties that generalize standard depth properties postulated for Euclidean data are established for the metric halfspace depth. The depth median, defined as the deepest point, is shown to have high robustness as a location descriptor in theory and simulation. An efficient algorithm is proposed to approximate the metric halfspace depth and illustrate its ability to adapt to the intrinsic data geometry. The metric halfspace depth was applied to an Alzheimer's disease study, revealing group differences in brain connectivity, modelled as covariance matrices, for subjects in different stages of dementia.

EC267 Room 709 HIGH-DIMENSIONAL DATA ANALYSIS Chair: Alexander Petersen

E0258: **Constrained approaches in learning high-dimensional sparse structures: Statistical optimality and optimization tools**
Presenter: **Shihao Wu**, University of Michigan, Ann Arbor, China

Sparse structures are ubiquitous in high-dimensional statistical models. To learn sparse structures from data, penalized approaches have been widely used and studied in the past two decades. Constrained approaches, however, were understudied due to their computational intractability and have recently been regaining attention given algorithmic advances in the optimization community and hardware improvements. Constrained approaches with penalized approaches in feature selection in high-dimensional sparse linear regression are compared. Specifically, it is focused on false discovery in the early stage of the solution path, which tracks how features enter and leave the model for a selection approach. As a penalized approach, LASSO is known to suffer false discoveries in the early stage. It is shown that best subset selection (BSS), as a constrained approach, achieves exact zero false discovery throughout the early stage under an optimal condition that it is found. It is referred to as zero false discovery as sure early selection. A solution to BSS within a tolerable optimization error suffices shown to achieve sure early selection. Extensive numerical experiments also demonstrate the advantages of constrained approaches on the solution path over penalized methods. Results for high-dimensional supervised fusion and high-dimensional mediation analysis and their corresponding constrained approaches will also be introduced and discussed.

E0988: **Computational strategies for regression model selection in the high-dimensional case**
Presenter: **Marios Demosthenous**, National Technical University of Athens, Greece
Co-authors: Cristian Gatu, Erricos Kontoghiorghes

Computational strategies for finding the best-subset regression models are proposed. The case of high-dimensional (HD) data where the number of variables exceeds the number of observations is considered. Within this context, a theoretical combinatorial solution is proposed. It is based on a regression tree structure that generates all possible subset models. An efficient branch-and-bound algorithm that finds the best submodels without generating the entire tree is adapted to the HD case. Furthermore, the R package lmSubsets is employed in the HD case to identify the best submodel based on the AIC family selection criteria. Preliminary experimental results are presented and analyzed. The efficient extension of the lmSelect algorithm to HD is discussed.

E1141: **Forward variable selection for ultra-high dimensional models**
Presenter: **Toshio Honda**, Hitotsubashi University, Japan

Forward variable selection procedures are described with stopping rules for feature screening in ultra-high dimensional quantile regression models and ultra-high dimensional generalized varying coefficient models. For such very large models, penalized methods like Lasso and SCAD do not work numerically, and some preliminary feature screening is necessary before such penalized methods are applied. The desirable theoretical properties of the forward procedures are presented by taking care of uniformity w.r.t. subsets of covariates properly. The necessity of such uniformity has been often overlooked in the literature. The stopping rules suitably incorporate the model size at each stage of the forward variable selection procedures. The results of numerical studies are also presented, and it is talked about possible extensions.

Thursday 03.08.2023 07:30 - 09:10 Parallel Session K – EcoSta2023

EV269 Room 704 ESTIMATION AND INFERENCE (VIRTUAL) Chair: Daoji Li

E0436: **Doubly robust identification and estimation of the LATE model with a continuous treatment**
Presenter: **Yingying Dong**, University of California Irvine, United States

Identification and estimation of the LATE model with a continuous treatment are considered. Two alternative restrictions are discussed on the first-stage instrument effect heterogeneity that allows for causal identification - monotonicity and treatment rank similarity. The former is popular in the LATE literature, while a slightly stronger version of the latter is exploited in the non-separable IV model literature. Neither assumption implies the other. Both assumptions can, at best, be partially tested. Causal estimands with doubly robust properties are proposed in that they are valid under either alternative restrictions.Further semiparametric estimators are proposed, and the asymptotic properties of these estimators are derived. When monotonicity holds, the primary estimand reduces to the standard LATE Wald ratio; otherwise, when treatment rank similarity holds, the approach allows for identifying treatment effect heterogeneity at different (conditional) treatment quantiles. The proposed estimators are applied to evaluate the impacts of neighbourhood poverty rate (a continuous treatment variable) on adults' labour market outcomes using the Moving to Opportunity (MTO) social experiment.

E1299: **Inference for low-rank models without rank estimation**
Presenter: **Hyukjun Kwon**, Rutgers University, United States
Co-authors: Yuan Liao, Jungjun Choi

A new debiasing procedure for linear low-rank models is introduced, where the parameter of interest is a high-dimensional matrix coefficient. The procedure achieves asymptotic normality without requiring knowledge of the true rank of the parameter matrix. The key feature of this approach is the use of diversified weights. An intermediate estimator is projected onto low-rank linear spaces that are estimated using these weights. Notably, this projection is robust to the rank misspecification. However, the estimated projection matrices are inconsistent with the true projections, creating new challenges in characterizing the asymptotic distribution of the debiased estimator. Nonetheless, the proposed debiasing procedure successfully addresses these issues. Lastly, the procedure does not require sample splitting.

E0264: **Inverse weighted quantile regression with partially interval-censored data**
Presenter: **Yeji Kim**, Korea university, Korea, South
Co-authors: Sangbum Choi, Seohyeon Park, Dipankar Bandyopadhyay, Taehwa Choi

A new inverse-probability censoring weighted (IPCW) estimating procedure for censored quantile regression with partially interval-censored data that include doubly-censored (DC) data and partly interval-censored (PIC) data is proposed. In addition to a certain amount of exact observations, DC data have either left-censored or right-censored data. In contrast, PIC data contain some interval-censored data, frequently occurring in the medical registry or HIV/AIDS clinical studies. Although various complex estimating techniques have been developed for censored quantile regression with DC and PIC data, a more simple and intuitive IPCW adjustment is considered, which can be effectively implemented by assigning a proper inverse-probability weight to each subject with an exact failure time observation. Asymptotic properties, including uniform consistency and weak convergence, are established for the resulting estimators. Further, an augmented-IPCW (AIPCW) estimation approach is discussed to gain more efficiency. Moreover, the proposed method can be readily adapted to handle multivariate partially interval-censored data. Simulation studies show the excellent finite-sample performance of the new inference procedure. The practical utility of the method is illustrated by an analysis of progression-free survival data from a phase III metastatic colorectal cancer clinical trial.

E0931: **A semiparametric approach in estimating sample maximum distribution**
Presenter: **Taku Moriyama**, Yokohama City University, Japan

A semiparametric approach is considered in estimating sample maximum distribution in iid settings. The sample maximum distribution is approximated by the generalized extreme value distribution. However, the convergence rate of the fitting estimator heavily depends on the tail index and gets slow as the index tends to zero. First, a fully nonparametric approach is introduced as an alternative approach and reports its asymptotic properties. The convergence rate of the nonparametric estimator with the optimal regularization parameter does not depend on the tail index under regularity conditions. Hence, the nonparametric approach outperforms the fitting estimator theoretically and numerically for distributions with the tail index around zero. On the other hand, the numerical accuracy of the nonparametric estimator is very poor for distributions with a tail index far from zero. A semiparametric mixture of the two approaches is proposed to develop a new approach complementing each other. Cross-validation and maximum-likelihood methods are provided for the mixing ratio selection, and reduction in computational cost is also discussed. The simulation experiment results and discusses the semiparametric estimator's numerical properties estimator are discussed.

EO125 Room 03 MODERN STATISTICAL INFERENCE FOR COMPLEX DATA (VIRTUAL) Chair: Meimei Liu

E0398: **Fitting low-rank models on egocentrically sampled partial networks**
Presenter: **Tianxi Li**, University of Virginia, United States

The statistical modelling of random networks has been widely used to uncover complex system interaction mechanisms and predict unobserved links in real-world networks. In many applications, network connections are collected via egocentric sampling: a subset of nodes is sampled first, after which all links involving this subset are recorded; all other information is missing. Compared with the assumption of "uniformly missing at random", egocentrically sampled partial networks require specially designed modelling strategies. Current statistical methods are either computationally infeasible or based on intuitive designs without theoretical justification. Here, an approach is proposed to fit general low-rank models for egocentrically sampled networks, which include several popular network models. This method is based on graph spectral properties and is computationally efficient for large-scale networks. It results in consistent recovery of missing subnetworks due to egocentric sampling for sparse networks. To our knowledge, this method offers the first theoretical guarantee for egocentric partial network estimation in the scope of low-rank models. The technique is evaluated on several synthetic and real-world networks, and it is shown that it delivers competitive performance in link prediction tasks.

E1038: **Operator-induced structural variable selection at scale: iBART and materials GWAS**
Presenter: **Meng Li**, Rice University, United States

In the emerging field of materials informatics, a fundamental task is to identify physicochemically meaningful descriptors, or materials genes, which are engineered from primary features and a set of elementary algebraic operators through compositions. Such materials genome-wide association studies, or materials GWAS, pose unprecedented challenges to statistical analysis partly due to the astronomically large number of correlated predictors with limited sample size. This problem is formulated as a variable selection with operator-induced structure (OIS), and a new method is proposed to achieve unconventional dimension reduction by utilizing the geometry embedded in OIS. Although the model remains linear, nonparametric variable selection for effective dimension reduction is iterated. This enables variable selection based on ab initio primary features, leading to a method that is orders of magnitude faster than existing methods, with improved accuracy. To select the nonparametric module, a desired performance criterion is discussed that is uniquely induced by variable selection with OIS; in particular, a Bayesian Additive Regression Trees (BART)- based variable selection method, leading to iterative BART (iBART) is proposed to employ. Numerical studies show the superiority of the proposed method, which continues to exhibit robust performance when the input dimension is out of reach of existing methods. Applications to single-atom catalysis will be discussed.

E1078: **Multilayer network model for joint analysis of structural brain imaging vector and functional connectome matrix**
Presenter: **Shuo Chen**, University of Maryland, School of Medicine, United States

Assessing the association between brain structural imaging (SI) measures and functional connectome (FC) obtained from neuroimaging data is considered. In this network analysis, the outcomes are off-diagonal elements of an FC (covariance) matrix, while predictors are a multivariate vector of SI variables and other covariates. A multilayer network model is proposed to capture the systematic association patterns between subsets of SIs and FC sub-networks. The first layer network is a bipartite graph characterizing the association between all SI variables and FC outcomes, where an edge denotes a non-zero SI-FC association. A large proportion of edges are located within latent dense bipartite subgraphs, while other edges are randomly and sparsely distributed in the rest of the bipartite graph. The second layer network represents the connectomic graph, where most FC outcomes in the first layer dense subnetworks comprise dense clique subgraphs. The globally sparse and locally dense multilayer network model can reveal which FC subnetworks are systematically influenced by a selected subset of SIs. Algorithms are developed to identify the underlying multilayer sub-networks and propose a statistical inference framework to test these sub-networks. The approach is further applied to 4242 participants from UK Biobank to evaluate the effects of whole-brain white matter microstructure integrity and cortical thickness on the whole-brain FC network.

E1138: **Flow-based conditional predictive inference**
Presenter: **Xin Xing**, Virginia Tech University, United States
Co-authors: Youhui Ye, Meimei Liu

The objective of predictive inference is to determine precise levels of confidence in predictions for new objects using past experience. A novel method called Flow-based Conditional Predictive Inference (FCI) is introduced for building predictive sets for complex and high-dimensional data. FCI uses ideas from adversarial flow to transfer input data to a random vector with known distributions, allowing for the construction of a p-value for uncertainty quantification. Our approach is applicable and robust, even when the testing data is contaminated. The method, robust flow-based conformal inference, on benchmark datasets is evaluated, and it is demonstrated that it produces effective predictive sets and accurate outlier detection, outperforming other approaches in terms of power.

EO153 Room Virtual R01 INTERPRETABLE STATISTICS AND ML FOR BIOLOGICAL AND BIOMEDICAL DATA Chair: Wei Vivian Li

E0183: **Reference-informed spatial domain detection for spatial transcriptomics**
Presenter: **Xiang Zhou**, University of Michigan, United States

Spatial transcriptomics studies are becoming increasingly common and large, providing unprecedented opportunities for characterizing complex tissues' spatial and functional organization. A statistical method, IRIS, is presented that leverages single-cell RNA-seq (scRNA-seq) data to accurately and efficiently detect spatial domains on complex tissues in spatial transcriptomics. IRIS is capable of modelling multiple tissue slices jointly, explicitly accounts for the correlation both within and across slices, and takes advantage of numerous algorithmic innovations to achieve highly scalable computation. The advantages of IRIS through in-depth analysis of five spatial transcriptomics datasets from different technologies across distinct tissues and species are demonstrated. In real data applications, IRIS achieves 51% - 94% accuracy gain over existing methods in datasets with known ground truth. In addition, IRIS is faster than existing methods in moderate-sized datasets and is the only method applicable to large-scale spatial transcriptomics data collected today. As a result, IRIS captures the fine-scale structures of brain regions, reveals the spatial heterogeneity of tumour micro-environments, and characterizes the structural changes of the seminiferous tubes in the underlying testis diabetes, all at a speed and accuracy unattainable by existing approaches.

E0781: **Unsupervised learning approaches for bulk and single cell genomics**
Presenter: **Sushmita Roy**, University of Wisconsin-Madison, United States

Advances in genomic technologies have substantially expanded the repertoire of high-dimensional datasets that measure different types of modalities such as the transcriptome, epigenome and three-dimensional genome organization. As these datasets are sparse in addition to being high-dimensional, an open challenge is to effectively analyze these datasets to extract meaningful low-dimensional patterns that reflect interpretable cell, gene or region clusters. Non-negative Matrix Factorization (NMF) is a popular dimensionality reduction approach that has been used for diverse types of biological and non-biological datasets. In this talk, I will present extensions of NMF to tackle two problems in regulatory genomics. First, extensions of NMF for understanding the three-dimensional organization of the genome and its role in phenotypic variation. The results show that NMF is a powerful approach for analyzing 3D genome organization from Hi-C assays that can recover biologically meaningful topological units and also enable us to smooth sparse Hi-C datasets and also identify dynamics in 3D genome organization. In the second part of my talk, I will present applications of NMF and its extensions to analyze single-cell RNA-seq datasets. We will present an application of NMF for deriving robust cell clusters from scRNA-seq data of the developing hindbrain and spinal cord and how NMF can be extended to handle multi-sample data to identify common and context-specific cell clusters.

E0562: **Using community-wide data to address (some) challenges in single cell data**
Presenter: **Kim-Anh Le Cao**, University of Melbourne, Australia

Cell identity classification is an ongoing challenge for analysing single-cell RNA-seq (scRNA-seq) data. Numerous tools exist for predicting cell identity using single-cell reference atlases. However, many challenges remain, including correcting for inherent batch effects between reference and query data and insufficient phenotype data from the reference. The proposed method aims to build bulk transcriptome atlases as references against which single cell identity can be queried. The advantage is that bulk data often contain detailed phenotype information and that numerous high-quality bulk datasets can be reused. A new computational and statistical framework, Sincast (SINgle-cell data CASTing onto reference), will be introduced to project and query scRNA-seq data to bulk RNA-seq data using principal component analysis and diffusion map. Structural discrepancies between bulk and single-cell data are solved by either aggregating or imputing single cells and the most beneficial approach, depending on the data context, is discussed. Sincast can also be used to reveal intermediate single-cell states when projected against bulk data. The approach in several case studies is illustrated.

E0638: **Detection of short identity-by-descent segments using low-frequency variants**
Presenter: **Lu Zhang**, HKBU Institute for Research and Continuing Education, ShenZhen, Hong Kong

Identity by descent (IBD) segments are shared inherited nucleotide sequences from common ancestors that have applications as biomarkers during lineage analysis, disease mutation mapping, and within broader fields of population genetics. Current methods used to detect IBD segments focus on identifying long IBD segments, with short IBD segments (i.e., those with a length shorter than 2cM) often going undetected. SILO, a method with a remarkably improved ability, is presented to detect short IBD segments compared to state-of-the-art methods. SILO detects IBD segments by considering both common and low-frequency variants (LFV), where each LFV is assumed to follow a Bernoulli distribution with probability θ following a Beta distribution. The Beta distribution could be learned from population genomic variants. SILO was benchmarked against five methods (GERMLINE2, hap-IBD, HapFABIA, Parente, and TRUFFLE) on simulated data and a pedigree from the 1000 Genomes Project. The results show that SILO has an unrivalled ability to detect short IBD segments and that its ability to detect long IBD segments is similar to that of current methods in the field.

EO130 Room Virtual R02 MODERN AND INNOVATIVE STATISTICAL LEARNING METHODS FOR COMPLEX DATA Chair: Guannan Wang

E0246: TSSS: A novel triangulated spherical spline smoothing for data distributed on complex surfaces

Presenter: **Zhiling Gu**, Iowa State University, United States
Co-authors: Shan Yu, Guannan Wang, Lily Wang

Data distributed on surfaces has been widely observed and analyzed in practice, especially in Earth, planetary, and biomedical science. Examples include the estimation of geopotential for the Earth, predicting the magnetic North Pole's movement, and completing the cosmic microwave background radiation field. A novel nonparametric method is introduced to efficiently discover the underlying signals on surfaces of complex domains. In particular, a penalized spline estimator defined on a triangulation of surface patches with irregular shapes is proposed, which guarantees signal matching and smoothness. Moreover, the asymptotic behaviour of the proposed estimators is investigated, which indicates the proposed estimation enjoys a desirable convergence. Simulation experiments and data applications on cortical surface functional magnetic resonance imaging (cs-fMRI) data and oceanic near-surface atmospheric data are conducted, showing that the proposed method has advantages over existing methods.

E0455: Modeling and inference for 3D complex objects

Presenter: **Lily Wang**, George Mason University, United States
Co-authors: Guannan Wang, Yueying Wang

The use of 3D complex objects is growing in various applications as data collection techniques continue to evolve. Identifying and locating significant effects within these objects is essential for making informed decisions based on the data. An advanced nonparametric framework is presented for learning and inferring 3D complex objects, enabling accurate estimation of the underlying signals and efficient detection and localization of significant effects. The proposed method addresses the problem of analyzing 3D complex objects collected within irregular boundaries by modelling them as functional data and utilizing trivariate spline smoothing based on triangulations to estimate the mean functions. In addition, a novel approach is presented for constructing simultaneous confidence corridors to quantify estimation uncertainty, and the procedure is extended to accommodate comparisons between two independent samples. The proposed methods are illustrated through a real-data application using the Alzheimer's Disease Neuroimaging Initiative database.

E0473: Evaluating biomarkers for treatment selection from reproducibility studies

Presenter: **Xiao Song**, University of Georgia, United States
Co-authors: Kevin K Dobbin

Evaluating new or more accurately measured predictive biomarkers for treatment selection based on a previous clinical trial involving standard biomarkers is considered. Instead of rerunning the clinical trial with the new biomarkers, a more efficient approach requires only either conducting a reproducibility study in which the new biomarkers and standard biomarkers are both measured on a set of patient samples or adopting replicated measures of the error-contaminated standard biomarkers in the original study is proposed. This approach is easier to conduct and much less expensive than studies that require new samples from patients randomized to the intervention. In addition, it makes it possible to perform the estimation of the clinical performance quickly since there will be no requirement to wait for events to occur, as would be the case with prospective validation. The treatment selection is assessed via a working model, but the proposed estimator of the mean restricted lifetime is valid even if the working model is misspecified. The proposed approach is assessed through simulation studies and applied to a cancer study

E0553: Penalized deep partially linear cox models with application to CT scans of lung cancer patients

Presenter: **Yuming Sun**, University of Michigan, Ann Arbor, United States
Co-authors: Jian Kang, Chinmay Haridas, Nicholas Mayne, Alexandra Potter, Chi-Fu Jeffrey Yang, David Christiani, Yi Li

Lung cancer is a leading cause of cancer mortality globally, highlighting the importance of understanding its mortality risks to design effective patient-centred therapies. The National Lung Screening Trial (NLST) was a nationwide study aimed at investigating risk factors for lung cancer. The study employed computed tomography texture analysis (CTTA) to quantify the mortality risks of lung cancer patients. The challenge in identifying the texture features that impact cancer survival is due to their sensitivity to factors such as scanner type, segmentation, and organ motion. To overcome this challenge, a novel Penalized Deep Partially Linear Cox Model (Penalized DPLC) is proposed, which incorporates the SCAD penalty to select significant texture features and employs a deep neural network to estimate the nonparametric component of the model accurately. The convergence and asymptotic properties of the estimator are proved, and it is 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. Our findings provide valuable insights into the relationship between these factors and survival outcomes.

EO118 Room 201 HIGH-DIMENSIONAL MEDIATION ANALYSIS Chair: Shuoyang Wang

E1085: Comparisons of variable selection and inference methods in high-dimensional mediation analysis

Presenter: **Xizhen Cai**, Williams College, United States
Co-authors: Yeying Zhu, Yuan Huang

Mediation analysis is a framework to understand how a treatment affects the outcome through intermediate variables, namely mediators. Over the past decades, large and high-dimensional datasets have become easily stored and publicly available. This has led to many recent advances in mediation analysis, including developing models to fit more complex data structures and methods for mediator selections in high-dimensional settings. The statistical inference procedure following the mediator selection is also an important step in the mediation analysis. The effect of different variable selection and inference procedures is studied through simulation studies. The simulation settings and the findings are discussed to provide guidelines that help distinguish among various approaches, highlight the advantages and disadvantages of each, and identify ones that perform better in certain scenarios.

E1102: Quantile mediation analysis with convoluted confounding effects via deep neural networks

Presenter: **Shuoyang Wang**, Yale University, United States
Co-authors: Runze Li, Yuan Huang

Traditional mediation analysis methods have been limited to dealing with only a few mediators, and they face challenges when the number of mediators is high-dimensional. In practice, these challenges can be compounded by outliers and the complex relationships introduced by confounders. A novel quantile-based partially linear mediation analysis method (QMDNN) that can handle high-dimensional mediators is proposed to address these issues. Deep neural network techniques to model complex nonlinear relationships in confounders are introduced. Unlike most existing works focusing on mediator selection, estimation and inference on mediation effects are emphasized. Theoretical analysis shows that the proposed procedure controls type I error rates for hypothesis testing on mediation effects. When the dimension of mediators is high, the proposed method consistently selects important features in the outcome model. Numerical studies show that the proposed method outperforms existing approaches under a variety of settings, demonstrating the versatility and reliability of QMDNN as a modelling tool for complex data. The application of QMDNN to study DNA methylation's mediation effect of childhood trauma on cortisol stress reactivity reveals previously undiscovered relationships by providing a comprehensive profile of the relationship at various quantiles.

E1140: **A joint approach to screen high dimensional mediators in epigenetic data with repeated outcomes**
Presenter: **Yu Jiang**, University of Memphis, United States
Co-authors: Lu Xie, Hongmei Zhang, Meredith Ray, Cen Wu

There has been a growing demand for mediation analyses for high-dimensional data, specifically for high-dimensional epigenetic data, where the number of potential mediators is more than half a million. While existing statistical approaches conduct mediation analyses for single or multiple mediator models, none of these methods deals with high dimensional mediators while controlling for both Type I and Type II errors for repeated outcomes, which are often observed in longitudinal studies. There is no software or packages to perform an efficient screening. A novel screening method, screening, was developed to perform a screening process for high-dimensional mediators with a repeated outcome. Simulation studies were used to evaluate the performance of the proposed joint screening method. For both continuous and binary outcomes, the proposed joint screening method showed comparable sensitivity and specificity when the number of mediators in the model was relatively small and higher sensitivity when the number was larger compared to traditional FDR and Bonferroni methods. This proposed method was applied to real data to examine the mediation effects of DNA methylation in the association between maternal smoking and childhood asthma. It is a powerful tool for a better understanding the epigenetic effects of mediating risk factors on disease outcomes.

E1223: **Bayesian high-dimensional mediation analysis incorporating neighborhood information**
Presenter: **Yunju Im**, University of Nebraska Medical Center, United States

Mediation analysis is a useful tool to investigate the roles of the mediators that lie in the pathway from exposure to an outcome variable. With recent technological advances, researchers may encounter situations where the number of candidate mediators is high, necessitating mediator selection. In such cases, mediators may be correlated or exhibit network structures, making using ancillary information important in the mediator selection process. To address these challenges in a high-dimensional setting, a flexible statistical method is introduced that leverages external knowledge of the network structures between the mediators to improve selection and estimation accuracy. The proposed method's benefits are demonstrated through simulations and real-world data.

EO101 Room 203 CAUSAL INFERENCE: METHODS AND APPLICATIONS (VIRTUAL) Chair: Subir Ghosh

E0670: **An instrumental variable method for point processes: generalized Wald estimation based on deconvolution**
Presenter: **Shizhe Chen**, University of California, Davis, United States
Co-authors: Zhichao Jiang, Peng Ding

Point processes are probabilistic tools for modelling event data. While there is fast-growing literature studying the relationships between point processes, how such relationships connect to causal effects remains unexplored. In the presence of unmeasured confounders, parameters from point process models do not necessarily have causal interpretations. An instrumental variable method for causal inference is proposed with point process treatment and the outcome. Causal quantities based on potential outcomes are defined, and nonparametric identification results with a binary instrumental variable are established. The traditional Wald estimation is extended to deal with point process treatment and outcome, showing that it should be performed after a Fourier transform of the intention-to-treat effects on the treatment and outcome. Thus, it takes the form of deconvolution. This is termed the generalized Wald estimation, and an estimation strategy based on well-established deconvolution methods is proposed.

E0802: **A t-test for synthetic controls**
Presenter: **Yinchu Zhu**, Brandeis University, United States
Co-authors: Victor Chernozhukov, Kaspar Wuthrich

A practical and robust method is proposed for making inferences on average treatment effects estimated by synthetic controls. A K-fold cross-fitting procedure is developed for bias correction. To avoid the difficult estimation of the long-run variance, the inference is based on a self-normalized t-statistic, which has an asymptotically pivotal t-distribution. The t-test is easy to implement, provably robust against misspecification, valid with non-stationary data, and demonstrates excellent small-sample performance. Compared to difference-in-differences, the proposed method often yields more than 50% shorter confidence intervals and is robust to violations of parallel trends' assumptions. An R-package for implementing the methods is available.

E0822: **Covariate-adaptive randomization inference in matched designs**
Presenter: **Samuel Pimentel**, UC Berkeley, United States
Co-authors: Yaxuan Huang

It is common to conduct causal inference in matched observational studies by proceeding as though treatment assignments within matched sets are assigned uniformly at random and using this distribution as the basis for inference. This approach ignores observed discrepancies in matched sets that may be consequential for the distribution of treatment, which are succinctly captured by within-set differences in the propensity score. This problem is addressed via covariate-adaptive randomization inference, which modifies the permutation probabilities to vary with estimated propensity score discrepancies and avoids requirements to exclude matched pairs or model an outcome variable. It is shown that the test achieves type I error control arbitrarily close to the nominal level when large samples are available for propensity score estimation. The large-sample behaviour of the new randomization test for a difference-in-means estimator of a constant additive effect is characterized. It is also shown that existing sensitivity analysis methods generalize effectively to covariate-adaptive randomization inference. Finally, the empirical value of covariate-adaptive randomization procedures is evaluated via comparisons to traditional uniform inference in matched designs with and without propensity score callipers and regression adjustment using simulations and analysis of genetic damage among welders.

E0877: **Double-robust two-way-fixed-effects regression for panel data**
Presenter: **Lihua Lei**, Stanford University, United States

A new estimator is proposed for the average causal effects of a binary treatment with panel data in settings with general treatment patterns. The approach augments the two-way-fixed-effects specification with the unit-specific weights that arise from a model for the assignment mechanism. It is shown how to construct these weights in various settings, including situations where units opt into the treatment sequentially. The resulting estimator converges to an average (over units and time) treatment effect under the correct specification of the assignment model. It is shown that the estimator is more robust than the conventional two-way estimator: it remains consistent if either the assignment mechanism or the two-way regression model is correctly specified and performs better than the two- way-fixed-effect estimator if both are locally misspecified. This strong double robustness property quantifies the benefits of modelling the assignment process and motivates using our estimator in practice.

EO184 Room 503 BAYESIAN METHODS AND SCALABLE COMPUTATION FOR EMERGING STUDIES (VIRTUAL) Chair: Zhenke Wu

E0537: A Bayesian hierarchical model for mortality surveillance using partially verified verbal autopsy data

Presenter: **Zehang Li**, University of California, Santa Cruz, United States

Monitoring data on causes of death is an integral part of understanding the burden of diseases and evaluating public health interventions. Verbal autopsy (VA) is a well-established method for gathering information about deaths outside of hospitals by interviewing family members or caregivers of a deceased person. Data from VA can be used to infer causes of death based on the collected symptoms and covariates. However, little information about the relationship or dynamics between symptoms and the new cause of death is available when a new disease emerges. A Bayesian hierarchical model framework is proposed that can be used to estimate the fraction of deaths due to the emerging disease using the VA data stream collected with partially verified cause-of-death. A latent class model is used to capture the distribution of symptoms and their dependence parsimoniously. Several potential sources of bias are discussed that may occur in the data selection process of cause-of-death verification, and our framework is adapted to account for the verification mechanism. Also, structured priors are developed to improve prevalence estimation for sub-populations. Our model's performance is demonstrated using simulation and a mortality surveillance dataset that includes suspected COVID-19-related deaths in Brazil.

E0532: Tree-regularized Bayesian latent class analysis: Addressing weak separation in small-sized subpopulations

Presenter: **Zhenke Wu**, University of Michigan at Ann Arbor, United States

Latent class models (LCMs) have been used to derive dietary patterns, where class profiles represent a probability vector of exposures to a set of diet components queried on diet assessment tools. However, LCM-derived dietary patterns can exhibit strong similarities, or weak separation, resulting in unstable class profile estimates and less accurate class assignments. This issue is exacerbated in small-sized subpopulations. This issue is addressed with a newly proposed tree-regularized Bayesian LCM that shares statistical strength across dietary patterns. These patterns are guided by an unknown tree learned from the data to produce improved estimates of class profiles and assignments using limited data. This is achieved via a Dirichlet diffusion tree process that specifies a prior distribution for the unknown tree over classes. Dietary patterns that share proximity in the tree are shrunk towards ancestral dietary patterns a priori, with the degree of shrinkage varying across pre-specified food groups. Using dietary intake data from the Hispanic Community Health Study/Study of Latinos, the utility of our model is demonstrated to identify dietary patterns of US adults of South American ethnic backgrounds.

E0928: Bayesian methods for vaccine safety surveillance using federated data sources

Presenter: **Fan Bu**, UCLA, United States

A Bayesian sequential analysis framework for data sources distributed across a federated network motivated by vaccine safety surveillance studies is discussed. The purpose is to enable rapid detection of vaccine safety events from observational healthcare data that accrue over time. Our framework aims at resolving three main challenges: first, control of testing errors in sequential analyses of streaming data; second, correction of bias induced by observational data; third, distributed learning of federated data sources while preserving patient-level privacy. These challenges in a unified statistical framework are tackled by extracting profile likelihoods that retain rich distributional information while protecting individual-level data privacy and hierarchical analysis of adverse control outcomes. As evidenced by large-scale empirical evaluations using real-world data sources, the framework provides substantial improvements over existing approaches to safety surveillance.

E1267: Exact inference for stochastic epidemic models via uniformly ergodic block sampling

Presenter: **Raphael Morsomme**, Duke University, United States

Co-authors: Jason Xu

Stochastic epidemic models provide an interpretable probabilistic description of the spread of a disease through a population. Yet, fitting these models to partially observed data is a notoriously difficult task due to the intractability of the likelihood for many classical models. To remedy this issue, a novel data-augmented Markov chain Monte Carlo algorithm is introduced for exact Bayesian inference under the stochastic susceptible-infectious-removed model, given only discretely observed counts of infections. In a Metropolis-Hastings step, the latent data are jointly proposed from a surrogate process carefully designed to resemble the target process closely and from which epidemics consistent with the observed data can be efficiently generated. This yields a method that efficiently explores the high - dimensional latent space and easily scales to outbreaks with thousands of infections. Further, this Markov chain Monte Carlo algorithm is proved to be uniformly ergodic, and it is observed to mix much faster than existing single-site samplers. The algorithm is applied to fit a semi-Markov susceptible-infectious-removed model to the 2013-2015 outbreak of Ebola Haemorrhagic Fever in Gueckedou, Guinea.

EO233 Room 506 ADVANCES IN MODEL-BASED CLUSTERING Chair: Yingying Zhang

E0442: Model-based clustering for tensor-variate data

Presenter: **Salvatore Daniele Tomarchio**, University of Catania, Italy

Co-authors: Antonio Punzo, Luca Bagnato

More flexible statistical methodologies are necessary with the increasing complexity of real data. One type of data that exemplifies this need is tensor-variate (or multi-way) structures. However, real data often includes atypical observations that render the traditional normality assumption unsuitable. Two novel tensor-variate distributions that are heavy-tailed generalizations of the tensor-variate normal distribution are introduced to address this issue. These distributions are then used to construct finite mixture models for model-based clustering. The eigendecomposition of the components' scale matrices is utilized to reduce complexity in the models, resulting in two families of parsimonious tensor-variate mixture models. The parameter estimation employs variations of the EM algorithm. Since the number of parsimonious models is dependent on the order of the tensors, strategies are implemented to shorten the initialization and fitting processes. The effectiveness of these procedures is evaluated through simulated data analyses. Real data are also investigated.

E0840: On the estimation of multilevel cross-classified latent class models

Presenter: **Silvia Columbu**, University of Cagliari, Italy

Co-authors: Nicola Piras, Jeroen Vermunt

An extension of latent-class models is presented for dealing with the clustering of multilevel cross-classified data. The model is formulated to allow two levels of clustering, one for lower-level units and one for cross-classified units, i.e. observations simultaneously nested within two or more groups. Given the dependency structure in the data, maximum likelihood estimation cannot be directly performed using a standard EM algorithm. A variation including a stochastic step is proposed. Global model selection criteria are also provided to determine the number of latent classes at both levels of the multilevel structure. The performances of the estimation algorithm and the selection criteria are verified through simulation studies. An application to educational data is finally discussed.

E0603: **On model-based clustering of directional data with heavy tails and scatter**
Presenter: **Yingying Zhang**, Western Michigan Univesity, United States
Co-authors: Volodymyr Melnykov, Igor Melnykov

Directional statistics deals with data that can be naturally expressed in the form of vector directions. Von Mises-Fisher distribution is one of the most fundamental parametric models to describe directional data. Mixtures of von Mises-Fisher distributions represent a popular approach to handling heterogeneous populations. However, such models can be affected by the presence of noise, outliers, and heavy tails. To relax these model limitations, a mixture of contaminated von Mises-Fisher distributions is proposed. The performance of the proposed methodology is tested on synthetic data and applied to the data containing abstracts from the Joint Statistical Meetings held in Denver in 2008. The obtained results demonstrate the importance of the proposed procedure and its superiority over the traditional mixture of von Mises-Fisher distributions in the cases of heavy tails or scatter.

E0637: **Model-based clustering on the spatial-temporal and intensity patterns of tornadoes**
Presenter: **Rong Zheng**, Western Illinois University, United States
Co-authors: Yana Melnykov, Yingying Zhang

Tornadoes are one of nature's most violent windstorms that can occur all over the world except Antarctica. Previous scientific efforts were spent studying this natural hazard from genesis, dynamics, detection, forecasting, warning, measuring, and assessing. At the same time, the aim was to model the tornado datasets using modern, sophisticated statistical and computational techniques. The goal is to develop novel finite mixture models and perform cluster analysis on tornadoes' spatial-temporal and intensity patterns. First, to analyze the tornado dataset, a Gaussian distribution with the mean vector and variance-covariance matrix represented is used as exponential functions of intensity and time. Then, a Gaussian mixture model is employed, with mean vector and variance-covariance represented as exponential functions of intensity and time. Thirdly, manly transform parameters are added to the Gaussian mixture model to take care of the skewness in the tornado dataset. Computer algorithms obtain results. A summary of insights is provided about tornado forecasting and assessing.

E0775: **Transformation mixture modeling for skewed data groups with heavy tails and scatter**
Presenter: **Yana Melnykov**, The University of Alabama, United States
Co-authors: Volodymyr Melnykov, Xuwen Zhu

For decades, Gaussian mixture models have been the most popular mixtures in literature. However, the adequacy of the fit provided by Gaussian components is often in question. Various distributions capable of modelling skewness or heavy tails have been considered in this context recently. A novel contaminated transformation mixture model is proposed that is constructed based on the idea of transformation to symmetry and can account for skewness and heavy tails and automatically assign scatter to secondary components.

EO182 Room 603 RECENT ADVANCES IN NOWCASTING (VIRTUAL) Chair: Ekaterina Smetanina

E0327: **Panel data nowcasting: The Case of price-earnings ratios**
Presenter: **Eric Ghysels**, University of North Carolina Chapel Hill, United States
Co-authors: Andrii Babii, Jonas Striaukas, Ryan Ball

The proposed method uses structured machine learning regressions for nowcasting with panel data consisting of series sampled at different frequencies. Motivated by the problem of predicting corporate earnings for a large cross-section of firms with macroeconomic, financial, and news time series sampled at different frequencies, the method focuses on the sparse-group LASSO regularization, which can take advantage of the mixed-frequency time series panel data structures. The empirical results show the superior performance of the machine learning panel data regression models over analysts' predictions, forecast combinations, firm-specific time series regression models, and standard machine learning methods.

E0770: **Back to the present: Learning about the Euro Area through a now-casting model**
Presenter: **Michele Modugno**, Federal Reserve Board, United States
Co-authors: Danilo Cascaldi-Garcia, Thiago Ferreira

A multi-country model is built for simultaneously now-casting economic conditions in the euro area and its three largest member countries, Germany, France, and Italy. The model formalizes how market participants and policymakers monitor in real-time both euro-area and country-specific market moving indicators. The out-of-sample evaluation corroborates the usefulness of a multi-country approach to monitor the euro area. Indeed, the model provides real-time accurate predictions of economic conditions both on average and in the past three recessions while finding that soft data is timely and intrinsically informative.

E0801: **Nowcasting recession risk in the US and the Euro area**
Presenter: **Domenico Giannone**, University of Washington, United States
Co-authors: Francesco Furno

Timely coincident recession risk indicators are presented for the United States (US) and the Euro Area (EA) at a monthly frequency. The indicators are constructed by estimating a parsimonious Bayesian logit based on two predictors, which summarize financial conditions and real economic activity. The Composite Indicator of Systemic Stress (CISS) is selected to measure financial conditions, the US PMIs, and the EA Economic Sentiment Index (ESI) to summarize real economic activity. These predictors are available immediately after the month of reference concludes. Back-testing the indicator over the periods 1980-2021 for the US and 1985-2021 for the Euro Area reveals a 96% and a 92% in-sample accuracy and 95% and 88% pseudo-out-of-sample, respectively. The indicators are more accurate than popular indicators such as the Sahm-Rule - especially at determining when the economy leaves a recession - and complement spread-based indicators, which are good at forecasting instead of nowcasting recessions.

EO031 Room 604 RECENT ADVANCES IN HIGH DIMENSIONAL TIME SERIES (VIRTUAL) Chair: Danna Zhang

E0539: **Robust estimation of high dimensional time series**
Presenter: **Danna Zhang**, University of California, San Diego, United States

High dimensional non-Gaussian time series data are increasingly encountered in various applications. It makes many traditional statistical analysis tools for independent data infeasible and poses a great challenge in developing new tools for time series. A novel Bernstein-type inequality for high-dimensional time series shall be presented. Then it is applied to investigate two high-dimensional robust estimation problems: (1) time series regression with fat-tailed and correlated covariates and errors, (2) fat-tailed vector autoregression. As a natural requirement of consistency, the dimension can be allowed to increase exponentially with the sample size under a very mild moment and dependence conditions.

E0785: **CP factor model for dynamic tensors**
Presenter: **Yuefeng Han**, University of Notre Dame, United States

Observations in various applications are frequently represented as a time series of multidimensional arrays, called tensor time series, preserving the inherent multidimensional structure. A factor model approach is presented, in a form similar to tensor CP decomposition, to the analysis of high-dimensional dynamic tensor time series. As the loading vectors are uniquely defined but not necessarily orthogonal, it is significantly different from the existing tensor factor models based on Tucker-type tensor decomposition. The model structure allows for a set of uncorrelated one-dimensional latent dynamic factor processes, making it much more convenient to study the underlying dynamics of the time series. A new high-order projection estimator is proposed for such a factor model, utilizing the special structure and the idea of the higher-order orthogonal iteration procedures commonly used in the Tucker-type tensor factor model and general tensor CP decomposition procedures. Theoretic al investigation provides statistical error bounds for the proposed methods, which shows the significant advantage of utilizing the special model structure.

E0798: **Statistical inference of spectral density for high dimensional time series**
Presenter: **Chi Zhang**, University of California, San Diego, United States
Co-authors: Danna Zhang

Spectral density plays a fundamental role in time series analysis. There has been a well-developed asymptotic theory for the spectral estimates in the low-dimensional case. For high-dimensional time series, distributional theory on spectral density is still lacking. This paper aims to establish an inference theory on high dimensional spectral density. In particular, a Gaussian approximation result is established for the maximum deviation of the spectral density estimate over frequencies, which can be used to tackle a variety of time series inference problems. Furthermore, two different resampling methods are introduced to implement high-dimensional spectral inference in practice and provide theoretical justification for their validity.

EO163 Room 605 RECENT ADVANCES IN STOCHASTIC MODELLING (VIRTUAL) Chair: Takashi Owada

E1113: **Quickest detection of the change of community via stochastic block models**
Presenter: **Ruizhi Zhang**, University of Georgia, United States

Community detection is a fundamental problem in network analysis and has important applications in sensor networks and social networks. In many cases, the network's community structure may change at some unknown time. Thus, it is desirable to come up with efficient monitoring procedures that can detect the change as quickly as possible. The Erdős-Rényi model and the bisection stochastic block model (SBM) are used to model the pre-change and post-change distributions of the network, respectively. Initially, it is assumed there is no community in the network. However, at some unknown time, a change occurs, and two communities are formed in the network. An efficient monitoring procedure is proposed using the number of k-cycles in the graph. The asymptotic detection properties of the proposed procedure are derived when all parameters are known. A generalized likelihood ratio (GLR) type detection procedure and an adaptive CUSUM type detection procedure are constructed to address the problem when parameters are unknown.

E1216: **Bayesian inference for causal effects under interference with a partially observed diffusion process on networks**
Presenter: **Fei Fang**, Yale University, United States
Co-authors: Laura Forastiere, Edoardo Airoldi, Amir Ghasemianlangroodi

Behaviours are likely to spread in a connected population, and the presence of a behavioural intervention may boost this spread. The setting is considered where it is observed at baseline, the set of treated units, and at baseline and follow-up the social network and the prevalence of behaviours. To investigate the problem, a network-based diffusion model is assumed, including the network susceptible-infected-susceptible (SIS) model and network susceptible-infected (SI) model, formulated as a continuous-time Markov process. A Bayesian data augmentation procedure is developed to impute over time the behavioural change as a result of diffusion from social ties or as a result of the intervention for the treated. Based on the estimated parameters, an imputation method is used to evaluate the causal effects of hypothetical treatment allocations with different rates and network-based strategies. Under simplified network models, closed forms were also derived for the expected effect of increasing the treatment rate under different baseline behaviour prevalence and network structures. The proposed method is applied to a factorial randomized experiment delivering a behavioural intervention in villages in Honduras under different treatment rates and strategies. This data allows us to compare adoption rates under a hypothetical strategy imputed in one arm with the actual adoption rates observed in the arm assigned to that strategy.

E1279: **Large deviations for the volume of k-nearest neighbor balls**
Presenter: **Takashi Owada**, Purdue University, United States
Co-authors: Christian Hirsch, Taegyu Kang

The large deviations theory is developed for the point process associated with the Euclidean volume of k-nearest neighbour balls centred around the points of a homogeneous Poisson or a binomial point process in the unit cube. Two different types of large deviation behaviours of such point processes are investigated. The first result is the Donsker-Varadhan large deviation principle, under the assumption that the centring terms for the volume of k-nearest neighbour balls grow to infinity more slowly than those needed for Poisson convergence. Additionally, large deviations are also studied based on the notion of M0-topology, which takes place when the centring terms tend to infinitely sufficiently fast, compared to those for Poisson convergence. As applications of the main theorems, large deviations are discussed for the number of Poisson or binomial points of degree at most k in a random geometric graph in the dense regime.

E1305: **Optimal transport-based domain adaptation for sensor data with application in smart manufacturing**
Presenter: **Rui Xie**, Universify of Central Florida, United States
Co-authors: Dazhong Wu

Tool wear prediction plays a crucial role in smart manufacturing, where advanced technologies and data-driven insights are employed to optimize production processes and minimize downtime caused by tool failures. In recent years, numerous machine learning-based predictive modeling approaches have emerged for tool wear prediction. However, accurately predicting tool wear under varying operating conditions, such as depth of cut, feed rate, and workpiece material, remains a daunting task due to the intricate nature of tool wear mechanisms. To tackle this challenge head-on, an algorithm leveraging optimal transport (OT)-based domain adaptation has been developed. This innovative approach enables the transfer of knowledge on tool wear from one operating condition to another. The effectiveness of the OT-based transfer learning model has been verified using a limited dataset encompassing diverse operating conditions. Through rigorous experimentation, the results have demonstrated a substantial improvement in tool wear prediction accuracy achieved by the OT-based transfer learning method.

EO029 Room 606 RECENT ADVANCES IN TIME SERIES TREND AND CHANGE POINT ANALYSIS (VIRTUAL) Chair: Kin Wai Chan

E1184: **Recursive nonparametric estimation: Principles, methods and applications**
Presenter: **Man Fung Leung**, University of Illinois Urbana-Champaign, United States
Co-authors: Kin Wai Chan

Existing long-run variance estimators face a dilemma between mean squared error, time complexity, and space complexity. A conceptual decomposition will be presented to understand this phenomenon. The new insights allow us to improve existing works, but further efficient estimators in a principle-driven way are characterized. The asymptotic theory and simulations show that this new approach leads to online estimators with a lower mean squared error. It is also discussed practical enhancements such as mini-batch and automatic updates. Encouraging finite-sample results are illustrated in online change point detection, stochastic approximation, and Markov chain Monte Carlo convergence diagnosis.

E1236: **Estimation of the long-run error variance in nonparametric regression with time series errors**
Presenter: **Marina Khismatullina**, Erasmus University Rotterdam, Netherlands
Co-authors: Michael Vogt

A new difference-based estimator of the long-run error variance for nonparametric regression is proposed in the case that the error terms have an autoregressive structure. Such an estimator is required for virtually all inferential procedures in the context of nonparametric regression. The proposed estimator improves on existing methods in several respects. First, the estimator produces accurate estimation results even when the AR process is quite persistent. Second, it produces accurate results even in the presence of a very pronounced regression function. These properties are illustrated by a simulation study that compares the proposed estimator with existing ones.

E1271: **A general framework for constructing locally self-normalized multiple-change-point tests**
Presenter: **Cheuk Hin Cheng**, The Chinese University of Hong Kong, Hong Kong
Co-authors: Kin Wai Chan

A general framework is proposed to construct self-normalized multiple-change-point tests with time series data. The only building block is a user-specified single-change-detecting statistic, which covers a large class of popular methods, including the cumulative sum process, outlier-robust rank statistics, and order statistics. The proposed test statistic does not require robust and consistent estimation of nuisance parameters, selection of bandwidth parameters, or pre-specification of the number of change points. The finite-sample performance shows that the proposed test is size accurate, robust against misspecification of the alternative hypothesis, and more powerful than existing methods. A case study of the Shanghai-Hong Kong Stock Connect turnover is provided.

E1254: **Tight-difference-based centrosymmetric kernel estimators for long-run variance**
Presenter: **Kin Wai Chan**, The Chinese University of Hong Kong, Hong Kong
Co-authors: Ya Xuan Wang

Long-run variance estimation is important for many statistical inference procedures. Existing methods give disappointing results when time series data exhibit both time-varying mean trends and significant serial dependence. Differencing and the kernel method are mainstream methods for resolving these problems and achieving mean robustness and consistency, respectively. Nevertheless, a large differencing lag is required to make kernel variance estimators work for time series, but the de-trending effect is substantially affected. It makes mean-robustness and consistency hard to be achieved simultaneously. This problem is tackled by constructing a novel centrosymmetric kernel to apply a tight differencing operation. The centrosymmetrization principle is simple and applicable to a large class of kernels. Optimal tight differencing sequences for handling serially dependent data are proven to be data-independent and universal. Therefore, they can be implemented directly without fitting into practice. The proposed principlealso appliese to spectral density estimation, where the robustness property is well inherited, showing great discrimination between the spectral pattern within noises and signals. This research was partially supported by General Research Fund 14304420, 14306421, and 14307922 provided by the Research Grants Council of HKSAR.

EO157 Room 702 EFFICIENT METHODS FOR FITTING COMPLEX NETWORK MODELS (VIRTUAL) Chair: Can Minh Le

E0644: **High order joint embedding for multi level link prediction**
Presenter: **Yubai Yuan**, Penn State University, United States

Link prediction infers potential links from observed networks and is one of the essential problems in network analyses. In contrast to traditional graph representation modelling, which only predicts two-way pairwise relations, a novel tensor-based joint network embedding approach is proposed on simultaneously encoding pairwise links and hyperlinks onto a latent space, which captures the dependency between pairwise and multi-way links in inferring potential unobserved hyperlinks. The major advantage of the proposed embedding procedure is that it incorporates both the pairwise relationships and subgroup-wise structure among nodes to capture richer network information. In addition, the proposed method introduces a hierarchical dependency among links to infer potential hyperlinks and leads to better link prediction. Theoretically, the estimation consistency is established for the proposed embedding approach and provides a faster convergence rate than link prediction utilizing pairwise links or hyperlinks only. Numerical studies on both simulation settings and Facebook ego networks indicate that the proposed method improves both a hyperlink and pairwise link prediction accuracy compared to existing link prediction algorithms.

E0789: **Efficient online reinforcement learning policies for continuous environments**
Presenter: **Mohamad Kazem Shirani Faradonbeh**, University, United States

One of the most popular dynamical models for continuous environments is linear systems that evolve according to stochastic differential equations. An interesting problem in this class of systems is learning to design control actions to minimize a quadratic cost function when system matrices are unknown. Implementable online reinforcement learning policies that learn the optimal control actions fast are discussed. In fact, the proposed policy efficiently balances exploration versus exploitation by carefully randomizing the parameter estimates such that the regret grows as the square root of time multiplied by the number of parameters. Theoretical performance analysis and simulations for learning to control an aeroplane will be presented to show efficiency.

E0800: **Approximate sampling and estimation of partition functions using neural networks**
Presenter: **George Cantwell**, Santa Fe Institute, United States

The closely related problems of sampling from a distribution known up are considered to be a normalizing constant and estimating said normalizing constant. The purpose is to show how variational autoencoders (VAEs) can be applied to this task. VAEs are trained to fit data drawn from an intractable distribution in their standard applications. The logic and train of the VAE are inverted to fit a simple and tractable distribution on the assumption of a complex and intractable latent distribution specified up to normalization. This procedure constructs approximations without the use of training data or Markov chain Monte Carlo sampling. The method on three examples are illustrated: the Ising model, graph clustering, and ranking.

E0910: **Joint spectral clustering in multilayer degree-corrected stochastic blockmodels**
Presenter: **Jesus Arroyo**, Texas A&M University, United States
Co-authors: Joshua Agterberg, Zachary Lubberts

Modern network datasets often have multiple layers, either as different views, time-varying observations, or independent sample units. These data require models and methods that are flexible enough to capture local and global differences across the networks while simultaneously being parsimonious and tractable to yield computationally efficient and theoretically sound solutions capable of aggregating information across the networks. The multilayer degree-corrected stochastic blockmodel is considered where a collection of networks share the same community structure, but degree-corrections and block connection probability matrices are permitted to be different. The identifiability of this model is established, and a spectral clustering algorithm is proposed for community detection in this setting. The theoretical results demonstrate that the misclustering error rate of the algorithm improves exponentially with multiple network realizations, even in the presence of significant layer heterogeneity. Simulation studies show this approach improves existing multilayer community detection methods in this challenging regime. Furthermore, in a case study of US airport data from January 2016 - September 2021, it is found that this methodology identifies meaningful community structure and trends in airport popularity influenced by pandemic impacts on travel

EO037 Room 705 RECENT ADVANCES IN HIGH-DIMENSIONAL STATISTICS AND MACHINE LEARNING (VIRTUAL) Chair: Xiucai Ding

E0380: **CLT for LSS of unnormalized sample covariance matrices when the dimension is much larger than the sample size**
Presenter: **Zhenggang Wang**, UC Davis, China
Co-authors: Xiucai Ding

Consider sample covariance matrices of the form $Q = \Sigma^{1/2}XX^{*}\Sigma^{1/2}$, where X is an $M \times N$ random matrix whose entries are independent random variables with mean zero and variance $1/\sqrt{NM}$, and Σ is a deterministic positive definite diagonal $M \times M$ matrix. The linear eigenvalue statistics of Q in the regime when the dimension M is much larger than the sample size N. The divergence of M/N would result in diverging support of such unnormalized sample covariance matrices. Contrary to some existing literature which normalized the matrices and approximated their spectral distribution by the semicircle law, it is shown that the MP law in this regime will still yield good results for unnormalized matrices. In particular, the anisotropic local law is established for the unnormalized matrices, and a central limit theorem is proved for the linear spectral statistics of Q both in the macroscopic and mesoscopic regime. Moreover, explicit formulas for the mean and covariance functions propose statistical applications in several different areas.

E0606: **Empirical Bayes estimation: When does g-modelling beat f-modelling in theory (and in practice)?**
Presenter: **Yandi Shen**, University of Chicago, United States
Co-authors: Yihong Wu

Empirical Bayes (EB) is a popular framework for the large-scale inference that aims to find data-driven estimators to compete with the Bayesian oracle that knows the truth prior. Two principled approaches to EB estimation have emerged over the years: f-modelling, which constructs an approximate Bayes rule by estimating the marginal distribution of the data, and g-modelling, which estimates the prior data and then applies the learned Bayes rule. For the Poisson model, the prototypical examples are the celebrated Robbins estimator and the nonparametric MLE (NPMLE), respectively. It has long been recognized in practice that while being conceptually appealing and computationally simple, Robbin's estimator lacks robustness and can be easily derailed by "outliers" (data points rarely observed before). A theoretical justification for the superiority of NPMLE over Robbins for heavy-tailed data is provided by considering priors with bounded pth moment previously studied for the Gaussian model. For the Poisson model with sample size n, assuming p>1 (for otherwise triviality arises), it is shown that the NPMLE with appropriate regularization and truncation achieves a total regret $O(n^{3/(2p+1)})$, which is minimax optimal within logarithmic factors. In contrast, the total regret of Robbin's estimator (with similar truncation) is $O(n^{3/(p+2)})$ and hence suboptimal by a polynomial factor.

E0744: **Spiked tensor model**
Presenter: **Jiaoyang Huang**, University of Pennsylvania, United States

The spiked tensor model is discussed, where one needs to extract information from a noisy high-dimensional data tensor. The algorithmic aspect of this model will be considered. First, the tensor power iteration algorithm will be considered, a natural generalization of the matrix power iteration. Necessary and sufficient conditions for the convergence of the power iteration algorithm are given. When the power iteration algorithm converges, for the rank one spiked tensor model, it is shown that the estimators for the spike strength and linear functionals of the signal are asymptotically Gaussian; for the multi-rank spiked tensor model, it is shown that the estimators are asymptotically mixtures of Gaussian. This new phenomenon is different from the spiked matrix model. Second, the tensor unfolding algorithm will be discussed. It results in a spiked random matrix where the number of rows (columns) grows polynomially in the number of columns (rows). By analyzing its spectrum, an exact threshold is obtained for the tensor unfolding algorithm, which is independent of the unfolding procedure. This threshold matches the conjectured computational threshold of the spiked tensor model.

E0902: **Contrastive learning: An expansion and shrinkage perspective**
Presenter: **Yiqiao Zhong**, UW Madison, United States
Co-authors: Cong Ma, Yu Gui

Contrastive learning is an unsupervised learning framework that has recently received much attention in the deep learning community. It achieves remarkable empirical performance, especially when no or few labelled training examples are available. Compared with the usual encoder-decoder structure in autoencoders, contrastive learning introduces positive/negative pairs and replaces the decoder with a projector. Intriguing puzzles about the role of projectors and dimensional collapse phenomena are often reported but not fully understood. Two major effects are identified, namely expansion and shrinkage, that contrastive learning promotes. The analysis is based on the Gaussian mixture model, which allows a systematic treatment despite its simplicity. The analysis reveals a rich phase transition phenomenon and characterizes generalization properties on downstream tasks, which closely match experimental results. The expansion and shrinkage perspective is a step toward demystifying the empirical puzzles and can potentially improve practice in self-supervised learning.

EO127 Room 708 RECENT ADVANCES IN FUNCTIONAL DATA ANALYSIS (VIRTUAL) Chair: Ruiyan Luo

E0555: **Continuous-time multivariate analysis: The transpose of functional data analysis**
Presenter: **Philip Reiss**, University of Haifa, Israel
Co-authors: Biplab Paul, Erjia Cui

The starting point for much multivariate analysis (MVA) is an $n \times p$ data matrix whose n rows represent observations and whose p columns represent variables. But some multivariate data sets may be best conceptualized not as n discrete p-variate observations but as p curves or functions defined on a common time interval. This viewpoint may be useful for data observed at very high time resolution, with unequal time intervals, and/or with substantial missingness. A framework for extending techniques of MVA is introduced to such settings by representing the curves as linear combinations of basis functions such as B-splines. This is formally identical to the Ramsay-Silverman representation of functional data. Still, whereas functional data analysis extends MVA to the case of observations that are curves rather than vectors heuristically, $n \times p$ data with p infinite is instead concerned with what happens when n is infinite. A simulation study demonstrates that the proposed continuous-time approach can improve the estimation of correlations among time series. A new R package, "ctmva", that translates the classical MVA methods of principal component analysis, Fisher's linear discriminant analysis, and k-means clustering is demonstrated in the above continuous-time setting. The methods are illustrated with a novel perspective on the well-known Canadian weather data set and with applications to neurobiological and environmental metric data.

E0377: **Functional degradation modeling of battery lives**
Presenter: **Pang Du**, Virginia Tech, United States
Co-authors: Quyen Do, Yili Hong

Renewable energy is critical for combating climate change, whose first step is the storage of electricity generated from renewable energy sources. Li-ion batteries are a popular kind of storage unit. Their continuous usage through charge-discharge cycles eventually leads to degradation. This can be visualized by plotting voltage discharge curves (VDCs) over discharge cycles. Studies of battery degradation have mostly concentrated on modelling degradation through one scalar measurement summarizing each VDC. Such simplification of curves can lead to inaccurate predictive models. The degradation of rechargeable Li-ion batteries from a NASA data set through modelling and predicting their full VDCs are analyzed. With techniques from longitudinal and functional data analysis, a new two-step predictive modelling procedure is proposed for functional responses residing on heterogeneous domains. The shapes are first predicted, and the domain end points of VDCs using functional regression models. Then these predictions are integrated to perform a degradation analysis. The approach is fully functional, allows the incorporation of usage information, produces predictions in a curve form, and thus provides flexibility in assessing battery degradation. Through extensive simulation studies and cross-validated data analysis, the approach demonstrates better prediction than the existing approach of modelling degradation directly with aggregated data.

E0303: **General nonlinear function-on-function regression via functional universal approximation**
Presenter: **Ruiyan Luo**, Georgia State University, United States

Various function-on-function (FOF) regression models with certain forms have been proposed to study the relationship between functional variables. However, because functional variables take values in infinite-dimensional spaces, the relationships between them can be much more complicated than those between scalar variables. The forms in existing FOF models may not be enough to cover a wide variety of relationships between functional variables, and hence the applicability of these models can be limited. A general nonlinear FOF regression model without any specific assumption on the model form is considered. To fit the model, inspired by the universal approximation theorem for the neural networks with "arbitrary width", a functional universal approximation theorem is developed, which asserts that a wide range of general maps between functional variables can be approximated with arbitrary accuracy by members in the proposed family of maps. This family is "full" functional in that the complexity of maps within the family is completely determined by the smoothness of coefficient functions in the map. With this functional universal approximation theorem, a novel method is developed to fit the general nonlinear FOF regression model, which includes all existing FOF models as special cases. The complexity of the fitted model is controlled by smoothness regularization without the necessity to choose the number of hidden layers or hidden neurons.

EO248 Room 709 ADVANCES IN DIMENSION REDUCTION: THEORY AND APPLICATIONS (VIRTUAL) Chair: Wenhui Sheng

E0349: **Central quantile subspace and its extension to functional data**
Presenter: **Eliana Christou**, University of North Carolina at Charlotte, United States

Quantile regression (QR) is becoming increasingly popular due to its relevance in many scientific investigations. There is a great amount of work on linear and nonlinear QR models. Specifically, the nonparametric estimation of conditional quantiles received particular attention due to its model flexibility. However, nonparametric QR techniques are limited in the number of covariates. Dimension reduction offers a solution to this problem by considering low-dimensional smoothing without specifying any parametric or nonparametric regression relation. The existing dimension reduction techniques focus on the entire conditional distribution. On the other hand, attention is turned to dimension-reduction techniques for conditional quantiles. A new method is introduced for reducing the dimension of predictor X. The methodology's performance is demonstrated through simulation examples and data applications, especially for financial data. Finally, an extension to functional data is presented, including an application on fMRI data.

E0512: **Pivot statistics for normal populations**
Presenter: **Liqiang Ni**, University of Central Florida, United States

Pivot statistics are widely used as the basis for statistical inference, e.g. confidence interval, hypothesis testing, and predictive inference. The concepts of sufficiency, pivot, invariance, and Bayesian inference are introduced. Then, the linkage between them is discussed, and some surprising results are shown: some classic statistic long thought to be pivot is not pivot; reducing the requirement of affine invariance to low-triangular invariance can produce a pivot. Some work-in-progress in the search for "true" pivots are considered.

E0691: **Model-free feature screening for high-throughput semi-competing risks data with FDR control**
Presenter: **Chenlu Ke**, Virginia Commonwealth University, United States

Partial Identifying biomarkers that contribute to early detection and effective treatment of cancers is a vital yet ongoing research task, which is often characterized by high-throughput data generated in a massive and fast manner by omics technologies, along with complicated survival endpoints as cancer course often involves adverse events such as progression and recurrence. A new feature screening framework is proposed for high-throughput survival data subject to semi-competing risks. Compared with existing prototypes, the method does not require an estimation of the survival function and relaxes the common assumption of independent censoring. The sure screening property and the rank consistency property in the notion of sufficiency are established. A knockoff procedure is also developed for controlling false discoveries. The advantages of the proposed method are demonstrated by simulation studies and an application in discovering the prognostic significance of copy-number alterations in multiple Myeloma.

E0756: **Dimension reduction for tensor response regression models**
Presenter: **Chung Eun Lee**, Baruch College, United States
Co-authors: Xin Zhang, Lexin Li

A flexible model-free approach to the regression analysis of a tensor response and a vector predictor is proposed. Without specifying the specific form of the regression mean function, the estimation of the dimension reduction subspace that captures all the variations in the regression mean function is considered. A new nonparametric metric called tensor martingale difference divergence is proposed, and its statistical properties are studied. Built on this new metric, computationally efficient estimation and asymptotically valid procedures are developed. The method's efficacy through simulations and a real data application for e-commerce are demonstrated.

Thursday 03.08.2023 09:20 - 11:00 Parallel Session L – EcoSta2023

EO010 Room 02 LATENT VARIABLE MODELS AND APPLICATIONS Chair: Xiangbin Meng

E0831: **Variational Bayesian estimation in diagnostic classification models**
Presenter: **Kensuke Okada**, The University of Tokyo, Japan
Co-authors: Keiichiro Hijikata, Motonori Oka, Kazuhiro Yamaguchi

Recent developments in variational Bayesian estimation methods for diagnostic classification models (DCMs) are discussed. The widespread use of information technologies in educational environments has created a demand for cognitive diagnosis and personalized feedback based on data obtained from learning systems. However, data from these systems often include a large number of respondents, items, and attributes. To efficiently conduct cognitive diagnosis in such situations, variational Bayesian estimation methods have been developed for a large class of diagnostic classification models, including the deterministic input, noisy "and" gate (DINA) model, deterministic input, noisy "or" gate (DINO) model, multiple-choice DINA model, and saturated DCM. The proposed algorithm consists of iteratively repeating two steps, the variational E-step and the variational M-step, until convergence. This algorithm is an important component in the scalable estimation of the Q-matrix in the DINA model. To facilitate the application of the proposed methods and further methodological developments, an R package, variationalDCM, has been developed that implements the proposed estimation methods for DCMs. The developed algorithm and implementation provide an effective framework for the study and application of DCMs in modern educational environments.

E0171: **Network community detection using higher-order structures**
Presenter: **Ji Zhu**, University of Michigan, United States

Many real-world networks commonly exhibit an abundance of subgraphs or higher-order structures, such as triangles and by-fans, surpassing what is typically observed in randomly generated networks. However, statistical models accounting for this phenomenon are limited, especially when community structure is of interest. This limitation is coupled with a lack of community detection methods that leverage subgraphs or higher-order structures. A novel community detection method is proposed that effectively incorporates these higher-order structures within a network. A finite-sample error bound is also developed for community detection accuracy under an edge-dependent network model, including community and triangle structures. This error bound is characterized by the expected triangle degree, leading to the proposed method's consistency. To our knowledge, this is the first statistical error bound and consistency result considering a single network's community detection under a network model with dependent edges. Through simulations and a real-world data example, it is demonstrated that our method reveals network communities otherwise obscured by methods that disregard higher-order structures.

E0319: **MSAEM estimation for multidimensional four-parameter normal ogive models**
Presenter: **Xiangbin Meng**, Northeast Normal University, China

A mixed stochastic approximation expectation maximization (MSAEM) algorithm coupled with a Gibbs sampler is developed to compute the marginalized maximum a posteriori estimate (MMAPE) of a multidimensional four-parameter normal ogive (M4PNO) model. The proposed MSAEM algorithm has the computational advantages of the stochastic approximation expectation maximization (SAEM) algorithm for multidimensional data. It also alleviates the potential instability caused by label switching and improves the estimation accuracy. Simulation studies are conducted to illustrate the good performance of the proposed MSAEM method, where MSAEM consistently performs better than SAEM and some other existing methods in multidimensional item response theory. Moreover, the proposed method is applied to a real data set from the 2018 Programme for International Student Assessment (PISA) to demonstrate the usefulness of the 4PNO model and MSAEM in practice.

E0345: **Recent developments in variational inference algorithms for restricted latent class models**
Presenter: **Kazuhiro Yamaguchi**, University of Tsukuba, Japan

As diagnostic classification models, restricted latent class (RLC) models have been employed in the social sciences, especially in psychology or educational research. The RLC is a special case of a general latent class model or mixture model in which the latent variables are categorical, and the latent classes are used to classify individuals into understandable sub-populations. Information and communication technology provides a wealth of rich information sources for latent classes and enables extended RCL models. However, an increase in data sources can make it difficult to estimate the model parameter of complex RCL models. The Variational Bayesian (VB) inference method, employed for complex machine learning models, is a good choice for RLC models. It is a deterministic posterior approximation method that works faster than the Markov chain Monte Carlo method. The key to deriving the VB estimation algorithm for an RLC model is introducing an auxiliary variable to represent equality constraints on the general latent class models. These constraint variables provide tractable mean-field variational inference for the RCL models. Furthermore, the VB method for extended RCL models (such as hidden Markov RCL or two-level RCLs models) can be derived based on this formulation. Recent developments in VB inference for various types of RLC models were reviewed and discussed.

EO022 Room 03 EARLY PHASE CANCER CLINICAL TRIAL DESIGNS AND REPORTING GUIDELINES Chair: Yisheng Li

E0663: **Randomized phase I clinical trials in oncology**
Presenter: **Alexia Iasonos**, Memorial Sloan Kettering Cancer Center, United States
Co-authors: John OQuigley

The aims of Phase 1 trials in oncology have broadened considerably from simply demonstrating that the agent/regimen of interest is well tolerated in a relatively heterogeneous patient population to addressing multiple objectives under the heading of early-phase trials and, if possible, obtaining reliable evidence regarding clinical activity to lead to drug approvals via the Accelerated Approval approach or Breakthrough Therapy designation in cases where the tumours are rare, the prognosis is poor or where there might be an unmet therapeutic need. Constructing a Phase 1 design that can address multiple objectives within the context of a single trial is not simple. Randomisation can play an important role, but carrying out such randomisation according to the principles of equipoise is a significant challenge in the Phase 1 setting. Suppose the emerging data are not sufficient to address the aims early on definitively. In that case, a proper design can reduce biases, enhance interpretability, and maximise information so that the Phase 1 data can be more compelling. The aim is to outline objectives and design considerations that must be adhered to respect ethical and scientific principles required for research in human subjects in controlled, early-phase clinical trials.

E0431: **Local continual reassessment methods for dose finding and optimization in drug-combination trials**
Presenter: **Ruitao Lin**, The University of Texas MD Anderson Cancer Center, United States

Due to the limited sample size and large dose exploration space, obtaining a desirable dose combination is a challenging task in the early development of combination treatments for cancer patients. Most existing designs for optimizing the dose combination are model-based, requiring significant efforts to elicit parameters or prior distributions. Model-based designs also rely on intensive model calibration and may yield unstable performance in the case of model misspecification or sparse data. The aim is to propose employing local, under-parameterized models for dose exploration to reduce the hurdle of model calibration and enhance the design robustness. Building upon the framework of the partial ordering continual reassessment method (POCRM), local data-based CRM (LOCRM) designs are developed for identifying the maximum tolerated dose combination (MTDC), using toxicity only, and the optimal biological dose combination (OBDC), using both toxicity and efficacy, respectively. The LOCRM designs only model the local data from neighbouring dose combinations. Therefore, they are flexible in estimating the local space and circumventing unstable characterization of the entire dose-exploration surface. The simulation studies show that this approach has competitive performance compared to widely used methods for finding MTDC, and it has advantages over existing model-based methods for optimizing OBDC.

E1071: **Determining a follow-up period for cure rate estimation in an exploratory phase clinical trial**
Presenter: **Yumiko Ibi**, Kyoto University Hospital, Japan
Co-authors: Yuka Sano, Tosiya Sato, Kentaro Ueno, Takashi Omori

In the development of cancer immunotherapy, the number of long-term survivors who will never experience the event of interest and are considered cured is increasing. The cure rate, which is defined as the proportion of long-term survivors, may be an important metric for patients to make treatment decisions. Therefore, it is important to properly determine a follow-up period time for the cure rate estimation before proceeding to a confirmatory clinical trial. The purpose is to propose a method of determining the follow-up time in cancer clinical trials for the proper estimation of the cure rate and to evaluate the proposed method using the Kaplan-Meier estimator and its corrected estimator.

E1183: **Guidance on statistical items in the SPIRIT and CONSORT extensions for early phase dose-finding clinical trials**
Presenter: **Christina Yap**, The Institute of Cancer Research, United Kingdom

Early-phase dose-finding (EPDF) trials are vital for developing new interventions. They are typically phase I or I/II trials that use adaptive dose escalation/de-escalation strategies to determine a safe and potentially active dose range for subsequent trials. The quality of EPDF trial protocols and reports was notably variable and suboptimal. Consequently, the DEFINE (DosE FIndiNg Extensions) study has developed consensus-driven extensions for EPDF trial protocols (SPIRIT-DEFINE) and trial reports (CONSORT-DEFINE) based on the SPIRIT (Standard Protocol Items: Recommendations for Interventional Trials) 2013 and CONSORT (CONsolidated Standards Of Reporting Trials) 2010 guidelines. The focus is on both guidelines' statistical aspects of the new and modified items. Such items include a detailed elaboration of the trial design (e.g., adaptive features, the timing of interim analyses, planned dose range with starting dose(s), dose allocation method, interim decision-making criteria, expansion cohort(s), operating characteristics), clear definitions of analysis populations, and plans for handling intercurrent events. The importance of including specific items will be stressed, and practical advice will be offered with examples of how to address them. Implementing both guidelines will facilitate transparency, comprehensiveness and reproducibility of methods and reduce research inefficiencies in EDPF trials. Funder: UK MRC/NIHR. Presented on behalf of the DEFINE Group

EO071 Room 04 MODELING LONGITUDINAL AND TIME-TO-EVENT DATA: NEW DIRECTIONS AND INNOVATIONS Chair: Esra Kurum

E0240: **Understanding the dynamic impact of COVID-19 through competing risk modelling with bivariate varying coefficients**
Presenter: **Wenbo Wu**, New York University Grossman School of Medicine, United States

The COVID-19 pandemic has exerted a profound impact on patients with kidney failure. Motivated by request by the U.S. Centers for Medicare & Medicaid Services, the analysis of their post-discharge hospital readmissions and deaths in 2020 revealed that the COVID-19 effect has varied significantly with post-discharge time and time since the pandemic onset. However, the complex dynamics of the impact trajectories cannot be characterized by existing varying coefficient models. To address this issue, a bivariate varying coefficient model is proposed for competing risks within a cause-specific hazard framework, where tensor-product B-splines are used to estimate the surface of the COVID-19 effect. An efficient proximal Newton algorithm is developed to facilitate fitting the new model to the massive Medicare data for dialysis patients. Difference-based anisotropic penalization is introduced to mitigate model overfitting and the wiggliness of the estimated trajectories; various cross-validation methods are considered in determining optimal tuning parameters. Hypothesis testing procedures are designed to examine whether the COVID-19 effect varies significantly with postdischarge time and the time since the pandemic onset, either jointly or separately. Simulation experiments and applications for Medicare dialysis patients demonstrate the proposed methods' estimation accuracy, controlled type I error rate, sufficient statistical power, and real-world performance.

E0252: **Constrained multivariate functional principal components analysis for eye-tracking experiments**
Presenter: **Brian Kwan**, University of California, Los Angeles (UCLA), United States
Co-authors: Catherine Sugar, Damla Senturk

Social attention is a neural process relevant to biomarker research in autism spectrum disorder (ASD), and eye-tracking experiments offer extensive insights into attentional processes by providing gaze patterns to sensory stimuli. However, common analysis through summaries, such as total looking time durations in regions of interest, collapses data across trials. A novel multivariate functional outcome is proposed that carries looking time duration information from multiple regions of interest jointly as a function of trial type among static image and dynamic video trials in an Activity Monitoring task without collapsing across trials. A novel constrained multivariate functional principal components analysis is also proposed to capture variation in the proposed outcome, incorporating the constraint in the data that looking time durations from multiple regions of interest need, to sum up to the total trial time. Application to data from the Autism Biomarkers Consortium for Clinical Trials (ABC-CT) study yields new insights into dominant modes of variation of looking time durations from multiple regions of interest for school-age children with ASD and their typically developing peers.

E0254: **High-dimensional fixed effects profiling models and applications**
Presenter: **Danh Nguyen**, University of California, Irvine, United States

Profiling analysis aims to evaluate healthcare providers, such as hospitals, nursing homes, dialysis facilities, etc., concerning a patient's outcome. Fixed effects (FE) profiling methods have considered binary outcomes, such as 30-day hospital readmission or mortality. For the unique population of dialysis patients, (1) regular blood tests are required to evaluate the effectiveness of treatment and avoid adverse events, including dialysis inadequacy, imbalance mineral levels, and anaemia, among others, as well as (2) the need for continuous monitoring/care after transitioning to dialysis. The versatility of FE profiling models is illustrated through several applications in profiling dialysis facilities in the U.S. and recent FE model developments, including (a) time-varying/time-dynamic standardized readmission ratio, (b) profiling for recurrent adverse events, and (c) new insights on operating characteristics such performance of FE model under the low information context/sparse outcome data setting.

E0238: **A Bayesian multilevel time-varying framework for joint modelling of longitudinal and survival outcomes**
Presenter: **Esra Kurum**, University of California, Riverside, United States

Over 782,000 individuals in the U.S. have end-stage kidney disease, with about 72% of patients on dialysis, a life-sustaining treatment. Dialysis patients experience high mortality and frequent hospitalizations, at about twice per year. These poor outcomes are exacerbated at key time periods, such as the fragile period after the transition to dialysis. To study the time-varying effects of modifiable patient and dialysis facility risk factors on hospitalization and mortality, a novel Bayesian multilevel time-varying joint model is proposed. Efficient estimation and inference are achieved within the Bayesian framework using Markov Chain Monte Carlo, where multilevel (patient- and dialysis facility-level) varying coefficient functions are targeted via Bayesian P-splines. Applications to the United States Renal Data System, a national database which contains data on nearly all patients on dialysis in the U.S., highlight significant time-varying effects of patient- and facility-level risk factors on hospitalization risk and mortality. The finite sample performance of the proposed methodology is studied through simulations.

EO137 Room Virtual R01 RECENT ADVANCES IN STATISTCAL METHODS IN BIOMEDICAL APPLICATIONS Chair: Seung Jun Shin

E0411: **Evaluation of the natural history of disease by combining incident and prevalent cohorts: Application to the Nun Study**
Presenter: **Daewoo Pak**, Yonsei University, Korea, South

The Nun study is a well-known longitudinal epidemiology study of ageing and dementia that recruited elderly nuns who were not yet diagnosed with dementia (i.e., incident cohort) and had dementia prior to entry (i.e., prevalent cohort). In such a natural history of disease study, multistate modelling of the combined data from both incident and prevalent cohorts is desirable to improve inference efficiency. While important, the multistate modelling approaches for the combined data have been scarcely used in practice because prevalent samples do not provide the exact date of disease onset and do not represent the target population due to left truncation. It is demonstrated how to combine adequately both incident and prevalent cohorts to examine risk factors for every possible transition in studying the natural history of dementia. A four-state nonhomogeneous Markov model is adapted to characterize all transitions between different clinical stages, including plausible reversible transitions. The estimating procedure using the combined data leads to efficiency gains for every transition compared to those from the incident cohort data only.

E0422: **Envelope-based partial partial least squares with application to cytokine-based biomarker analysis for COVID-19**
Presenter: **Yeonhee Park**, University of Wisconsin, United States
Co-authors: Zhihua Su, Dongjun Chung

Partial least squares (PLS) regression is a popular alternative to ordinary least squares regression because of its superior prediction performance demonstrated in many cases. In various contemporary applications, the predictors include both continuous and categorical variables. A common PLS regression practice is treating the categorical variable as continuous. However, studies find that this practice may lead to biased estimates and invalid inferences. Based on a connection between the envelope model and PLS, an envelope-based partial PLS estimator is developed that considers the PLS regression on the conditional distributions of the response(s) and continuous predictors on the categorical predictors. Root-n consistency and asymptotic normality are established for this estimator. A numerical study shows that this approach can achieve more efficiency gains in estimation and produce better predictions. The method is applied for identifying cytokine-based biomarkers for COVID-19 patients, which reveals the association between the cytokine-based biomarkers and patients' clinical information, including disease status at admission and demographical characteristics. The efficient estimation leads to a clear scientific interpretation of the results.

E0619: **Scalable test of statistical significance for protein-DNA binding changes with insertion and deletion of bases**
Presenter: **Sunyoung Shin**, Pohang University of Science and Technology, Korea, South

Mutations in the noncoding DNA, which represents approximately 99% of the human genome, have been crucial to understanding disease mechanisms through the dysregulation of disease-associated genes. One key element in gene regulation that noncoding mutations mediate is the binding of proteins to DNA sequences. Insertion and deletion of bases (InDels) are the second most common type of mutations, following single nucleotide polymorphisms, that may impact protein-DNA binding. However, no existing methods can estimate and test the effects of InDels on the process of protein-DNA binding. A novel test of statistical significance, namely the binding change test (BC test), is developed using a Markov model to evaluate the impact and InDels altering protein-DNA binding is identified. The test predicts binding changer InDels of regulatory significance with an efficient importance sampling algorithm generating background sequences in favour of sizeable binding affinity changes. Simulation studies demonstrate its excellent performance. The application to human leukaemia data uncovers candidate pathological InDels on modulating MYC binding in leukemic patients. An R package atIndel is developed, which is available on GitHub.

E0718: **Transformed function on scalar regression for random distribution**
Presenter: **Hojin Yang**, Pusan National University, Korea, South

The aim is to develop a transformed function on the scalar regression model, using the functional principal components to account for random distribution. This framework allows us to model functions transformed from random distributions using the functional principal components approach in a transformed functional space and then regress functional principal component scores on multiple sets of predictors in their projected space. Thereby, the underlying model parameters, as well as the effect of the covariates in the projected space, can be estimated. Then, these parameters are transformed back to the original distributional space to understand the subject-specific random distributions. Hypothesis testing is also conducted, and predict random distributions for any given predictors are predicted.

EO159 Room Virtual R02 MODERN MACHINE LEARNING METHODS DEALING WITH A VARIETY OF DATA ISSUES Chair: Xinyi Li

E0456: **Nonparametric distributed learning of complex data**
Presenter: **Guannan Wang**, College of William & Mary, United States
Co-authors: Lily Wang, Shan Yu

Nowadays, one significant challenge in many applications comes from the enormous size of the datasets collected from modern technologies. To tackle this challenge, a novel nonparametric distributed learning method is designed based on multivariate spline smoothing over a triangulation of the domain. The proposed DHL algorithm has a simple, scalable, and communication-efficient implementation scheme that can almost achieve linear speedup. In addition, rigorous theoretical support is provided for the DHL framework. The DHL linear estimators have proven to reach the same convergence rate as the global spline estimators obtained using the entire dataset. The proposed DHL method is evaluated through extensive simulation studies and analyses of real applications.

E0991: **Fair conformal prediction**
Presenter: **Linjun Zhang**, Rutgers University, United States

Multi-calibration is a powerful and evolving concept originating in the field of algorithmic fairness. For a predictor f that estimates the outcome y given covariates x, and for a function class C, multi-calibration requires that the predictor $f(x)$ and outcome y are indistinguishable under the class of auditors in C. Fairness is captured by incorporating demographic subgroups into the class of functions C. Recent work has shown that, by enriching the class C to incorporate appropriate propensity re-weighting functions, multi-calibration also yields target-independent learning, wherein a model trained on a source domain performs well on unseen, future target domains(approximately) captured by the re-weightings. The multi-calibration notion is extended, and the power of an enriched class of mappings is explored. HappyMap, a generalization of multi-calibration, is proposed, which yields a wide range of new applications, including a new fairness notion for uncertainty quantification (conformal prediction), a novel technique for conformal prediction under covariate shift, and a different approach to analyzing missing data, while also yielding a unified understanding of several existing seemingly disparate algorithmic fairness notions and target-independent learning approaches. A single HappyMap meta-algorithm is given that captures all these results, together with a sufficiency condition for its success.

E1073: **Contrastive inverse regression for dimension reduction**
Presenter: **Didong Li**, University of North Carolina at Chapel Hill, United States

Supervised dimension reduction (SDR) has been a topic of growing interest in data science, as it enables the reduction of high-dimensional covariates while preserving the functional relation with certain response variables of interest. However, existing SDR methods are unsuitable for analyzing case-control study datasets. In this setting, the goal is to learn and exploit the low-dimensional structure unique to or enriched by the case group, also known as the foreground group. While some unsupervised techniques, such as the contrastive latent variable model and its variants, have been developed for this purpose, they fail to preserve the functional relationship between the dimension-reduced covariates and the response variable. A supervised dimension reduction method is proposed, called contrastive inverse regression (CIR), specifically designed for the contrastive setting. CIR introduces an optimization problem defined on the Stiefel manifold with a non-standard loss function. The convergence of CIR to a local optimum using a gradient descent-based algorithm and the numerical study empirically demonstrates the improved performance over competing methods for high-dimensional biomedical data are proved.

E1175: **Meta-analytic study of experimental data in the presence of missingness and unbalanced experimental factors**
Presenter: **Shyam Ranganathan**, Clemson University, United States
Co-authors: Raghupathy Karthikeyan, Qiong Su

Increased heat stress during cropping season poses significant challenges to rice production. A meta-analytic study of 1,946 experiments relating genetic factors (different rice species) was performed, and environmental factors (e.g., temperatures) to rice grain yield through random regression mixed models shows complex non-linear relationships. Rice yield is measured in terms of quantity components (panicle number, spikelet number per panicle, seed set rate, grain weight) and grain quality traits (milling yield, chalkiness, amylose, protein content). Model selection yields quadratic regression models,s and these models suggest both the optimum temperature ranges for high rice yields and the rice varieties most adaptive to temperature variations. However, there are significant variations across the large number of experiments used in the study due to missing data, variations in experimental conditions etc. While the number of experiments is large, each experiment includes few observations on a large number of variables, creating a high-dimensional problem. We propose a Bayesian framework to handle the meta-analysis in the presence of the limitations mentioned above. This will provide better results and help quantify the associated uncertainty. These results can be used to develop new field experiments to establish the relationships among phenotypic plasticity, genetic traits, and environmental factors.

EO115 Room 102 NEW DEVELOPMENTS IN IMAGING AND GENETICS USING LARGE SCALE STUDIES Chair: Haochang Shou

E0306: **Covariance-on-covariance regression**
Presenter: **Yi Zhao**, Indiana University, United States
Co-authors: Yize Zhao

A Covariance-on-Covariance regression model is introduced. It is assumed that there is (at least) a pair of linear projections on outcome covariance matrices and predictor covariance matrices such that a log-linear model links the variances in the projection spaces and additional covariates of interest. An ordinary least square type of estimator is proposed to simultaneously identify the projections and estimate model coefficients. Under regularity conditions, the proposed estimator is asymptotically consistent. The superior performance of the proposed approach over existing methods is demonstrated via simulation studies. Applying to data collected in the Human Connectome Project Aging study, the proposed approach identifies three pairs of brain networks, where functional connectivity within the resting-state network predicts functional connectivity within the corresponding task-state network. The three networks correspond to a global signal network, a task-related network, and a task-unrelated network. The findings are consistent with existing knowledge about brain function.

E0668: **Exploring cross-trait genetic architectures: Statistical models, computational challenges, and the BIGA platform**
Presenter: **Bingxin Zhao**, University of Pennsylvania, United States

Numerous statistical models have been proposed to analyze cross-trait genetic architectures utilizing summary statistics from genome-wide association studies (GWAS). However, systematically analyzing high-dimensional GWAS summary statistics presents logistical and computational challenges. The BIGA platform, a website, is introduced that offers unified data analysis pipelines and centralized data resources. A framework that implements statistical genetics tools on a cloud computing platform has been developed and integrated with extensive curated GWAS datasets. Furthermore, the recent theoretical analyses of the LD score regression (LDSC), a widely-used method for inferring heritability and genetic correlation using GWAS summary statistics, are discussed. The consistency and asymptotic normality of LDSC-based estimators are demonstrated, and the key factors that influence their performance are identified.

E0667: **Integration of fMRI and genomics data with interpretable multimodal deep learning**
Presenter: **Yu-Ping Wang**, Tulane University, United States

Integrating multi-modal brain imaging and genomics has been widely used in brain studies for improved diagnosis of mental diseases. Meanwhile, it calls for novel data integration models to capture complex associations between multi-modal brain imaging and genomics and, furthermore, interpretable approaches to uncover biological mechanisms with these models. An interpretable multi-modal integration model is developed to simultaneously perform automated disease diagnosis and result interpretation. It is named Grad-CAM-guided convolutional collaborative learning (gCAM-CCL) is achieved by combining intermediate feature maps with gradient-based weights. The gCAM-CCL model can generate interpretable activation maps to quantify pixel-level contributions of the input features. Moreover, the estimated activation maps are class-specific, which can therefore facilitate the identification of biomarkers underlying different groups. The gCAM-CCL model on a large cohort of brain imaging-genomics study is applied and validated, and its applications are demonstrated to both the classification of cognitive function groups and the discovery of underlying biological mechanisms.

E0805: **ME-Bayes SL: Enhanced Bayesian polygenic risk prediction leveraging information across multiple ancestry groups**
Presenter: **Jin Jin**, University of Pennsylvania, United States
Co-authors: Jianan Zhan, Jingning Zhang, Ruzhang Zhao, Jared O Connell, Yunxuan Jiang, 23andMe Research Team, Steven Buyske, Christopher Gignoux, Christopher Haiman, Eimear Kenny, Charles Kooperberg, Kari North, Bertram Koelsch, Genevieve Wojcik, Haoyu Zhang, Nilanjan Chatterjee

Polygenic risk scores (PRS) are now showing promising predictive performance on a wide variety of complex traits and diseases, but there exists a substantial performance gap across different populations. ME-Bayes SL, a method for the ancestry-specific polygenic prediction that borrows information in the summary statistics from genome-wide association studies (GWAS) across multiple ancestry groups, is proposed. ME-Bayes SL conducts Bayesian hierarchical modelling under a multivariate spike-and-slab model for effect-size distribution and incorporates an ensemble learning step to combine information across different tuning parameter settings and ancestry groups. ME-Bayes SL shows promising performance compared to alternatives in the simulation studies and data analyses of 16 traits across four distinct studies, totalling 5.7 million participants with substantial ancestral diversity. The method, for example, has an average gain in prediction R2 across 11 continuous traits of 40.2% and 49.3% compared to PRS-CSx and CT-SLEB, respectively, in the African Ancestry population. The best-performing method, however, varies by GWAS sample size, target ancestry, underlying trait architecture, and the choice of reference samples for LD estimation, and thus ultimately, a combination of methods may be needed to generate the most robust PRS across diverse populations.

EO113 Room 201 ECONOMETRICS AND STATISTICS ON UNOBSERVED HETEROGENEITY Chair: Katsumi Shimotsu

E0409: **Testing for unobserved heterogeneity in censored duration models: EM approach**
Presenter: **Katsumi Shimotsu**, University of Tokyo, Japan
Co-authors: Hiroyuki Kasahara, Hirokazu Matsuyama, Shota Takeishi

The method proposed to use a likelihood-based modified EM statistic to test for unobserved heterogeneity in censored duration models. The level and power of the modified EM test are compared with the likelihood ratio test (LRT) proposed by other researchers, an information matrix (IM) test constructed previously, and the Lagrange multiplier (LM) tests through a series of Monte Carlo simulations. The simulations show that the modified EM test outperforms the LRT, IM, and LM tests except when the sample size is no smaller than 200.

E0419: **A shrinkage likelihood ratio test for high-dimensional subgroup analysis with a logistic-normal mixture model**
Presenter: **Shota Takeishi**, University of Tokyo, Japan

The focus is on testing the existence of a subgroup with an enhanced treatment effect under the setting where high-dimensional covariates potentially characterize the subgroup membership. Using a logistic-normal mixture model, the method proposed a shrinkage likelihood ratio test built on a modified likelihood function that shrinks high-dimensional unidentified parameters towards zero when there exists no subgroup is proposed. This shrinkage helps handle the irregularity of the testing problem in the logistic-normal mixture model. It enables us to derive a tractable chi-squired-type asymptotic null distribution even under the high-dimensional regime. Simulation results will also be provided.

E0428: **Difference in differences with latent group structures**
Presenter: **Hiroyuki Kasahara**, University of British Columbia, Canada
Co-authors: Young Ahn

The identification of average treatment effects on the treated (ATT) is examined within latent group structures, where the distribution of potential outcomes depends on latent types. A scenario is explored in which parallel trends are maintained when conditioned on latent types but may not hold in aggregate, resulting in an inconsistent standard difference-in-difference estimator. It is demonstrated that the latent group-specific ATT (LGATT) can be identified when parallel trend assumptions and other regularity conditions are met for latent types. An estimator for the LGATT is proposed that minimizes a weighted least squares criterion function, using weights derived from the estimated posterior probabilities of each latent type.

E0602: **Identification and estimation of treatment effects in a linear factor model with fixed number of time periods**
Presenter: **Takuya Ishihara**, Tohoku University, Japan
Co-authors: Koki Fusejima

A new approach is provided for identifying and estimating the Average Treatment Effect on the Treated under a linear factor model that allows for multiple time-varying unobservables. Unlike the majority of the literature on treatment effects in linear factor models, our approach does not require the number of pre-treatment periods to go to infinity to obtain a valid estimator. Our identification approach employs a certain nonlinear transformation of the time-invariant observed covariates that are sufficiently correlated with the unobserved variables. This relevance condition can be checked with the available data on pre-treatment periods by validating the correlation of the transformed covariates and the pre-treatment outcomes. Our identification approach provides an asymptotically unbiased estimator of the effect of participating in the treatment when there is only one treated unit, and the number of control units is large.

EO061 Room 203 CAUSAL MEDIATION ANALYSIS AND PRINCIPAL STRATIFICATION Chair: Caleb Miles

E1288: **Sensitivity analysis for principal ignorability violation in estimating complier and noncomplier average causal effects**
Presenter: **Trang Nguyen**, Johns Hopkins Bloomberg School of Public Health, United States

An important strategy for identifying principal causal effects, which are often used in settings with noncompliance, is to invoke the principal ignorability (PI) assumption. As PI is untestable, it is important to gauge how sensitive effect estimates are to its violation. We focus on the common one-sided noncompliance setting where there are two principal strata, compliers and noncompliers. Under PI, compliers and noncompliers share the same outcome-mean-given-covariates function under the control condition. For sensitivity analysis, we allow this function to differ between compilers and noncompliers in several ways, indexed by an odds ratio, a generalized odds ratio, a mean ratio, or a standardized mean difference sensitivity parameter. We tailor sensitivity analysis techniques (with any sensitivity parameter choice) to several types of PI-based main analysis methods, including outcome regression, influence function-based and weighting methods. We illustrate the proposed sensitivity analyses using several outcome types from the JOBS II study and provide code in the R-package PIsens.

E1314: **Estimands versus algorithms in studies with competing events and interest in treatment mechanism**
Presenter: **Takuya Kawahara**, Harvard Medical School and Harvard Pilgrim Health Care Institute, United States
Co-authors: Jessica Young

In the presence of competing events, investigators might be interested in a direct treatment effect on the event of interest that does not capture treatment effects on competing events. Classical survival analysis methods that treat competing events like censoring events, at best, converge to a controlled direct effect that captures the treatment effect under the complete elimination of competing events which is often difficult to imagine. Recently, separable direct effects were proposed, which are the effects of modified versions of the study treatment with mechanisms removed other than those directly affecting the event of interest. These alternative notions of direct effect may have more practical relevance. Examples of data-generating conditions will be illustrated under which controlled and separable direct effects are identified but may take different values to varying degrees, including possibly different signs. This provides insights into the degree to which using even an unbiased estimator for a controlled direct effect could misleadingly inform a separable direct effect when that is really of the underlying interest. Conditions are also considered under which neither effect is identified due to the presence of an unmeasured common cause of the event of interest and the competing event and the bias associated with consistent estimation algorithms for these different effects.

E0545: **Mediation in causal survival analysis under competing risks using longitudinal modified treatment policies**
Presenter: **Ivan Diaz**, NYU Langone Health, United States

Longitudinal modified treatment policies (LMTP) have been recently developed as a novel method to define and estimate causal parameters that depend on the natural value of treatment. LMTPs represent an important advancement in causal inference for longitudinal studies as they allow the non-parametric definition and estimation of the joint effect of multiple categorical, numerical, or continuous exposures measured at several time points. The LMTP methodology is extended to mediation problems with time-varying mediators. Identification results and non-parametric locally efficient estimators that use flexible data-adaptive regression techniques are presented to alleviate model misspecification bias while retaining important asymptotic properties such as root-n-consistency. An application in the estimation of the (positive) effect of intubation on survival amongst hospitalized COVID-19 patients is presented, and it is decomposed into its (negative) effect through acute kidney injury and its (positive) effect through other paths.

E1316: **Causal mediation with instrumental variables**
Presenter: **Kara Rudolph**, Columbia University, United States
Co-authors: Ivan Diaz, Nicholas Williams

Mediation analysis is a strategy for understanding the mechanisms by which interventions affect later outcomes. However, unobserved confounding concerns may be compounded in mediation analyses, as there may be unobserved exposure-outcome, exposure-mediator, and mediator-outcome confounders. Instrumental variables (IVs) are a popular identification strategy in the presence of unobserved confounding. However, in contrast to the rich literature on the use of IV methods to identify and estimate a total effect of a non-randomized exposure, there has been almost no research into using IV as an identification strategy to identify mediational indirect effects. In response, novel estimands are defined and nonparametrically identified -complier interventional direct and indirect effects—when two, possibly related, IVs are available, one for the exposure and another for the mediator. Nonparametric, robust, efficient estimators are proposed for these effects and related compiler natural direct and indirect effects, and they are applied to a housing voucher experiment.

EO210 Room 503 LARGE-SCALE BAYESIAN INFERENCE Chair: Mattias Villani

E0404: **Rates of convergence in Bayesian meta-learning**
Presenter: **Pierre Alquier**, ESSEC Business School, Singapore
Co-authors: Badr Eddine Cherief Abdellatif, Charles Riou

The rate of convergence of Bayesian learning algorithms is determined by two conditions: the behaviour of the loss function around the optimal parameter (Bernstein condition) and the probability mass given by the prior neighbourhoods of the optimal parameter. In meta-learning, multiple learning tasks are faced that are independent but are still expected to be related in some way. For example, the optimal parameters of all the tasks can be close to each other. It is then tempting to use the past tasks to build a better prior that we use to solve future tasks more efficiently. From a theoretical point of view, we hope to improve the prior mass condition in future tasks and, thus, the rate of convergence. It is proved that this is indeed the case. Interestingly, it is also proved that the optimal prior can be learned at a fast rate of convergence, regardless of the rate of convergence within the tasks (in other words, the Bernstein condition is always satisfied for learning the prior, even when it is not satisfied within tasks).

E0405: **Improving the accuracy of marginal approximations in likelihood-free inference via localisation**
Presenter: **David Nott**, National University of Singapore, Singapore
Co-authors: Christopher Drovandi, David Frazier

Likelihood-free methods are an essential tool for performing inference for implicit models which can be simulated but for which the corresponding likelihood is intractable. However, common likelihood-free methods do not scale well to a large number of model parameters. A promising approach to high-dimensional likelihood-free inference involves estimating low-dimensional marginal posteriors by conditioning only on summary statistics believed to be informative for the low-dimensional component and then combining the low-dimensional approximations in some way. It is demonstrated that such low-dimensional approximations can be surprisingly poor in practice for seemingly intuitive summary statistic choices. Given an initial choice of low-dimensional summary statistic that might only be informative about a marginal posterior location, a new method which improves performance is described by first crudely localising the posterior approximation using all the summary statistics to ensure global identifiability, followed by a second step that hones in on an accurate low-dimensional approximation using the low-dimensional summary statistic. It is shown that the posterior this approach targets can be represented as a logarithmic pool of posterior distributions based on the low-dimensional and full summary statistics, respectively.

E0955: **High-dimensional conditionally Gaussian state space models with missing data**
Presenter: **Joshua Chan**, Purdue University, United States
Co-authors: Aubrey Poon, Dan Zhu

An efficient sampling approach is developed for handling complex missing data patterns and a large number of missing observations in conditionally Gaussian state space models. Two important examples are dynamic factor models with unbalanced datasets and large Bayesian VARs with variables in multiple frequencies. A key insight underlying the proposed approach is that the joint distribution of the missing data conditional on the observed data is Gaussian. Moreover, this conditional distribution's inverse covariance or precision matrix is sparse, and this special structure can be exploited to speed up computations substantially. The methodology is illustrated using two empirical applications. The first application combines quarterly, monthly and weekly data using a large Bayesian VAR to produce weekly GDP estimates. The second application extracts latent factors from unbalanced datasets involving over a hundred monthly variables via a dynamic factor model with stochastic volatility.

E1224: **Flexible variational Bayes based on a copula of a mixture of normals**
Presenter: **Robert Kohn**, University of New South Wales, Australia
Co-authors: David Nott, David Gunawan

Variational Bayes methods approximate the posterior density by a family of tractable distributions and use optimisation to estimate the unknown parameters of the approximation. The variational approximation is useful when exact inference is intractable or very costly. A flexible variational approximation is developed based on a copula of a mixture of normals, which is implemented using the natural gradient and a variance reduction method. The efficacy of the approach is illustrated by using simulated and real datasets to approximate multimodal, skewed and heavy-tailed posterior distributions, including an application to Bayesian deep feedforward neural network regression models. Each example shows that the proposed variational approximation is much more accurate than the corresponding Gaussian copula and a mixture of normal variational approximations.

EO131 Room 506 COMPLEX SPACE-TIME STRUCTURES WITH MODERN SPATIO-TEMPORAL STATISTICS Chair: Abhi Datta

E0209: **Classes of multivariate and space-time power-law covariance functions**
Presenter: **Pulong Ma**, Iowa State University, United States

Understanding marginal covariance and cross-covariance structures is essential for modelling continuously indexed multivariate and space-time processes. The Matern covariance function with short-range dependence has enabled several notable developments for multivariate and space-time models in the past few decades. However, many geophysical methods possess long-range dependence in space and space-time domains, which the Matern-based covariance models often fail to capture. The purpose is to exploit a scale-mixture framework to address this issue to construct new classes of multivariate and space-time covariance functions with power-law decay in the tail. This framework generates a covariance function by mixing over a base covariance model with a probability measure. Sufficient and necessary conditions are established to characterize the relationship between the behaviour of resultant covariance functions and the mixing probability measure. Then theoretical properties of the resultant covariance models are investigated. Several validity conditions that ensure the positive definiteness of the proposed multivariate covariance models are derived. The interplay among long-range dependence, Markov property, and screening effect are examined theoretically. Extensive simulation examples and real datasets are used to illustrate the superior performance of the proposed covariance models over the state-of-the-art models.

E0480: **Neural Bayes estimators for fast and efficient inference with spatial peaks-over-threshold models**
Presenter: **Jordan Richards**, King Abdullah University of Science and Technology, Saudi Arabia
Co-authors: Matthew Sainsbury-Dale, Andrew Zammit Mangion, Raphael Huser

Likelihood-based inference for spatial extremal dependence models is often infeasible in moderate or high dimensions due to an intractable likelihood function and/or the need for computationally-expensive censoring to reduce estimation bias. Neural Bayes estimators are a promising recent approach to inference that use neural networks to transform data into parameter estimates. They are likelihood free, inherit the optimality properties of Bayes estimators, and are substantially faster than classical methods. Neural Bayes estimators are adapted for peaks-over-threshold dependence models; in particular, a methodology is developed for coping with the computational challenges often encountered when modelling spatial extremes (e.g., censoring). It is demonstrated substantial improvements in computational and statistical efficiency relative to conventional likelihood-based approaches using popular extremal dependence models, including max-stable, and r-Pareto, processes and random scale mixture models.

E0536: **Extract long-term trend from large-missing gap Atlantic meridional overturning circulation (AMOC) data**
Presenter: **Yizi Cheng**, University of Cincinnati, United States
Co-authors: Won Chang, Roman Olson, Jongsoo Shin, Soon-il An

Atlantic Meridional Overturning Circulation (AMOC) is a large-scale ocean circulation that transports cold and dense water in the deep North Atlantic to the equator and warm and salty water in the upper layers of the North Atlantic to the pole. AMOC plays an important role in heat and carbon transport, thus having considerable impacts on climate change, and in response, on natural and human systems. There is an assumption about persistent AMOC weakening in response to anthropogenic forcing. However, due to technical limitations, AMOC was barely measured before 2004, leading to the difficulty in understanding how AMOC has changed over the last century. The Denoising Variational Auto-encoder (DVAE) framework is applied to reconstruct the long-term trend of AMOC in the previous 140 years. A DAVE model is built based on an ensemble of CESM model runs, which can extract the long-term trend from noisy observations on AMOC strength and sea surface temperature data while filtering out the effects of natural variability and quantifying the estimation uncertainty. The trained model is then applied to real observations, the RAPID AMOC project, and NOAA Extended Reconstructed Sea Surface Temperature V5 data. The results show that there has been a persistent AMOC weakening over the last century, with a 0.997 probability of decreasing over 1SV.

E0773: **Multi-model ensemble analysis with neural network Gaussian processes**
Presenter: **Trevor Harris**, Texas A&M University, United States

Multi-model ensemble analysis integrates information from multiple climate models into a unified projection. However, existing integration approaches based on model averaging can dilute fine-scale spatial information and incur bias from rescaling low-resolution climate models. A statistical approach, called NN-GPR, is proposed using Gaussian process regression (GPR) with an infinitely wide deep neural network-based covariance function. NN-GPR requires no assumptions about the relationships between climate models, no interpolation to a common grid, and automatically downscales as part of its prediction algorithm. Model experiments show that NN-GPR can be highly skilful at surface temperature and precipitation forecasting by preserving geospatial signals at multiple scales and capturing inter-annual variability. The projections particularly show improved accuracy and uncertainty quantification skill in regions of high variability. This allows cheaply assessing tail behaviour at a 50 km spatial resolution without a regional climate model (RCM). Evaluations on reanalysis data and SSP2-4.5 forced climate models show that NN-GPR produces similar overall climatologies to the model ensemble while better capturing fine-scale spatial patterns. Finally, NN-GPR's regional predictions are compared against two RCMs, and it is shown that NN-GPR can rival the performance of RCMs using only global model data as input.

EO241 Room 603 RECENT ADVANCES IN FUNCTIONAL DATA/ LONGITUDINAL DATA Chair: Mengying You

E0901: **A projection-based diagnostic test for generalized functional regression models**
Presenter: **Hua Liang**, George Washington University, United States

A novel diagnostic test is proposed to check goodness-of-fit for generalized functional regression models. The proposed test is free of any distribution assumptions and can be used for various classical functional regression models. However, it is based on independence in distribution and hence includes mean-based and higher-order moment-based tests as special cases. The proposed test avoids any subjective selection of tuning parameters by integrating over the directions along which the functional variables project. As a result, it enhances the local power and overcomes the infinite dimensionality problem simultaneously. A rather simple implementation procedure is developed. The performance of the proposed test is evaluated through theory and extensive simulation studies. The proposed procedure is also applied to analyze Canadian Weather data and Chinese Air Pollution data, resulting in several interesting models which achieve higher interpretability and estimation accuracy than the existing methods.

E0677: **Dynamic hierarchical state space forecasting**
Presenter: **Ziyue Liu**, Indiana University School of Medicine, United States

Situations are considered when there are time series data from multiple units that share similar patterns when aligned in terms of an internal time. Internal time is defined as an index according to evolving features of interest. When mapped back to the calendar time, these time series can span different time intervals that can include the future calendar time of the targeted unit, over which the information can be borrowed from other units in forecasting the targeted unit. First, a hierarchical state space model is built for the multiple time series data in terms of the internal time, where the shared components capture the similarities among different units while allowing for unit-specific deviations. A conditional state space model is then constructed to incorporate the information of existing units as the prior information in forecasting the targeted unit. The information from the other units and the unit's history is incorporated by running the Kalman filtering based on the conditional state space model on the targeted unit. The forecasts are then transformed from internal time back into calendar time for ease of interpretation. A simulation study is conducted to evaluate the finite sample performance. Forecasting state-level new COVID-19 cases in the USA is used for illustration.

E1092: **Dynamic Poisson state space prediction model for automobile insurance**
Presenter: **Jiakun Jiang**, Beijing Normal University, China

Prediction modelling of claim frequency is important for pricing and risk management in nonlife insurance. It needs to be updated frequently with the insured population and technology changes. Existing methods are either done in an ad hoc fashion, such as parametric model calibration or less so for the purpose of prediction. A Dynamic Poisson state space (DPSS) model is developed, which can continuously update the parameters whenever new information becomes available. DPSS model allows for both time-varying and time-invariant coefficients. To account for smoothness trends of time-varying coefficients over time, smoothing splines are used to model time-varying coefficients. The smoothing parameters are objectively chosen by maximum likelihood. The model is updated using batch data accumulated at prespecified time intervals, which allows for a better approximation of the underlying Poisson density function. The simulation shows that the new model has significantly higher prediction accuracy compared to existing methods. This methodology has been applied to real-world automobile insurance claim data sets in China over six years. It demonstrates its superiority by comparing it with the results of competing models from the literature.

E0883: **Semiparametric bivariate hierarchical state space model with application to hormone circadian relationship**
Presenter: **Mengying You**, University of Pennsylvania, United States
Co-authors: Wensheng Guo

The adrenocorticotropic hormone and cortisol are critical in stress regulation and the sleep-wake cycle. Most research has focused on how the two hormones regulate each other regarding short-term pulses. Few studies have been conducted on the circadian relationship between the two hormones and how it differs between normal and abnormal groups. The circadian patterns are difficult to model as parametric functions. Directly extending univariate functional mixed effects models would result in a large dimensional problem and a challenging nonparametric inference. A semiparametric bivariate hierarchical state space model is proposed, in which a hierarchical state space model with a nonparametric population average and subject-specific components models each hormone profile. The bivariate relationship is constructed by concatenating two latent independent subject-specific random functions specified by a design matrix, leading to a parametric inference on the correlation. A computationally efficient state-space EM algorithm is proposed for estimation and inference. The proposed method is applied to a study of chronic fatigue syndrome and fibromyalgia. An erratic regulation pattern is discovered in the patient group in contrast to a circadian regulation pattern conforming to the day-night cycle in the control group.

EO133 Room 604 RECENT CONTRIBUTIONS TO NONPARAMETRIC AND SEMIPARAMETRIC MODELS Chair: Alexandra Soberon

E0759: **Efficient estimation of a semiparametric panel data model with common factors and spatial dependence: Testing ETS**
Presenter: **Antonio Musolesi**, University of Ferrara, Italy

International carbon markets are an appealing and increasingly popular tool for countries to regulate carbon emissions. By putting a price on carbon, carbon markets make pollution less attractive for regulated firms. However, many observers remain sceptical of initiatives such as the European Union Emissions Trading System (EUETS), whose price remained low (compared to the social cost of carbon). The aim is to shed light on this dilemma by analyzing the effect of the EU ETS on CO2 emissions with a semiparametric panel data model where several types of cross-sectional dependence (CSD) and heteroscedasticity are allowed. A new estimator that extends the commonly correlated effect (CCE) approach to this framework is proposed. However, the initial estimator ignores the CSD and heteroscedasticity, leading to a loss of efficiency. Thus Generalized Least Squares (GLS)-type estimators are proposed. Under rather standard conditions, the parametric estimators are shown to bepNT-consistent, and the asymptotic normality of the nonparametric estimators is also established. Further, the GLS-type estimators are shown to dominate the other. Monte Carlo experiments investigate small sample properties of the estimators, and an empirical application on the effect of the EU ETS is conducted.

E0177: **A projection based approach for interactive fixed effects panel data models**
Presenter: **Juan Manuel Rodriguez-Poo**, Universidad de Cantabria, Spain
Co-authors: Georg Keilbar, Alexandra Soberon, Weining Wang

The aim is to present a new approach to estimation and inference in panel data models with interactive fixed effects, where the unobserved factor loadings can be correlated with the regressors. A distinctive feature of the proposed approach is to assume a nonparametric specification for the factor loadings, which allows us to partially out the interactive effects using sieve basis functions to estimate the slope parameters directly. The new estimator adopts the well-known partial least squares form, and its consistency and asymptotic normality are shown. It is found that the limiting distribution of the estimator has a discontinuity when the variance of the tcharacteristictic parameter is near the boundaries, which makes the usual "plug-in" method used to estimate the asymptotic variance only valid pointwise and may produce either over- or under- coverage probabilities. It is shown that uniformity can be achieved by cross-sectional bootstrap. Later, the common factors are estimated using principal component analysis (PCA), and the corresponding convergence rates are obtained. A Monte Carlo study indicates good performance in terms of mean squared error. The methodology is applied to analyze the determinants of growth rates in OECD countries.

E0272: **Modelling intervals of minimum/maximum temperature in the Iberian Peninsula**
Presenter: **Vladimir Rodriguez-Caballero**, Instituto Tecnologico Autonomo De Mexico, Mexico
Co-authors: Esther Ruiz, Gloria Gonzalez-Rivera

The aim is to propose to model intervals of minimum/maximum temperatures observed at a given location by fitting unobserved component models to bivariate systems of centre and log-range temperatures. The centre and log-range temperature are decomposed into potentially stochastic trends and seasonal and transitory components. The method contributes to the debate on whether stochastic or deterministic components better represent the trend and seasonal components. The methodology is implemented to intervals of minimum/maximum temperatures observed monthly in four locations in the Iberian Peninsula, namely, Barcelona, Coru na, Madrid and Seville. The aim is to show that, at each location, the centre temperature can be represented by a smooth integrated random walk with a time-varying slope. At the same time, the log range seems better represented by a stochastic level. Centre and log-range temperatures are also shown to be unrelated. The methodology is then extended to model simultaneously minimum/maximum temperatures observed at several locations. A multi-level dynamic factor model is used to extract potential commonalities among centre (log-range) temperatures while allowing for heterogeneity in different areas. The model is fitted to intervals of minimum/maximum temperatures observed at a large number of locations in the Iberian Peninsula

E0543: **Testing for relevance of partially parametric models with parametric nulls**
Presenter: **Daniel Henderson**, University of Alabama, United States
Co-authors: Jiancheng Jiang

Tests of relevance are considered for partially parametric models. Specifically, it is tested that the entire nonparametric function is irrelevant in predicting the outcome variable. This results in parametric null models, which can be estimated via (non-linear) least-squares. The asymptotic theory is developed for our test statistics (both under the null and versus local alternatives), and valid bootstrap procedures are proposed for use in finite samples. Further testing for the relevance of a control function in simultaneous equation models (i.e., test for exogeneity) is considered. This requires us to extend our theory to account for generated regressors. Then a joint test for correct parametric specification and irrelevance is considered. Finally, an omnibus version of our test with improved power is considered. Simulations and an empirical exercise suggest that the tests perform well in finite samples.

EO106 Room 605 STATISTICAL MODELING OF RELATION DATA Chair: Tianxi Li

E0288: **Nonparametric inference on network effects of general relationship network data**
Presenter: **Wen Zhou**, Colorado State University, United States
Co-authors: Yuan Zhang, Wenqin Du

In recent years, the relationship network has received significant attention for its ability to provide unique insights into agent interactions across various fields. Most existing studies have primarily focused on modelling the association between the relationship network and other covariates using arguably restrictive parametric models while largely overlooking the inference of network effects, such as the reciprocal or sender-receiver effect. Testing network effects within a relationship network are particularly challenging due to edge dependence, which renders permutation-based methods inapplicable. The testing statistics utilize the reduced U-statistics and admit analytically tractable limiting distributions, overcoming the nontrivial sampling distributions of network moment statistics on relationship network effects caused by degeneration and indeterminacy of degeneracy order. The theoretical guarantee of the testing framework is established by investigating the Berry-Esseen bounds for the testing statistics. To showcase the practicality of the methods, two real-world relationship networks are applied to them, one in international trade and the other in faculty hiring networks.

E0356: **Fundamental limits of spectral clustering in stochastic block models**
Presenter: **Anderson Ye Zhang**, University of Pennsylvania, United States

A precise characterization of the performance of spectral clustering is given for community detection under Stochastic Block Models by carrying out sharp statistical analysis. Spectral clustering has an exponentially small error with matching upper and lower bounds with the same exponent, including the sharp leading constant. The fundamental limits established for the spectral clustering hold for networks with multiple and imbalanced communities and sparse networks with degrees far smaller than $\log n$. The key to the results is a novel truncated ℓ_2 perturbation analysis for eigenvectors and a new analysis idea of eigenvectors truncation.

E0495: **Nonparametric link prediction for networks and Bipartite graph**
Presenter: **Kehui Chen**, University of Pittsburgh, United States
Co-authors: Jiashen Lu

A nonparametric link prediction framework for networks and Bipartite graphs is discussed. In particular, it will be discussed how to understand the missing mechanism and to deal with missing observations, when and how to use side information for link prediction, and how to improve the prediction accuracy for new entries (nodes). The proposed statistical framework leads to a simple algorithm with competitive performance.

E0752: **Identification and estimation of network statistics with missing link data**
Presenter: **Matthew Thirkettle**, Rice University, United States

Informative bounds on network statistics are obtained in a partially observed network whose formation is explicitly modelled. Partially observed networks are commonplace due to, for example, partial sampling or incomplete responses in surveys. Network statistics (e.g., centrality measures) are not point identified when the network is partially observed. Worst-case bounds on network statistics can be obtained by letting all missing links take values zero and one. It is dramatically improved on the worst-case bounds by specifying a structural model for network formation. An important feature of the model is that it allows for positive externalities in the network-formation process. The network-formation model and network statistics are set identified due to the multiplicity of equilibria. A computationally tractable outer approximation of the joint identified region is provided for preferences determining network-formation processes and network statistics. A simulation study on Katz-Bonacich centrality found that worst-case bounds that do not use the network formation model are 44 times wider than the bounds from my procedure obtained.

EO310 Room 606 MULTIVARIATE PROBLEMS FOR STRUCTURED DEPENDENT DATA II Chair: Michal Pesta

E1019: **Weighted change-point tests for short-range dependent data**
Presenter: **Kata Vuk**, University of Regensburg, Germany
Co-authors: Herold Dehling, Martin Wendler

The focus is on non-parametric weighted change-point tests based on two-sample U-statistics. By a suitable choice of weights, one obtains tests that are able to detect changes in time series that occur very early or very late during the observation period. The limit distribution of those test statistics is investigated under the hypothesis that no change occurs but also under the alternative that there is a change in mean. To illustrate the results, simulations and applications to real-life data will be presented.

E1157: **Numerical inversion of characteristic functions for exact multivariate statistical inference**
Presenter: **Viktor Witkovsky**, Slovak Academy of Sciences, Slovakia

Computing the exact statistical distributions of multivariate test statistics is a challenging task. A method is proposed for calculating the distributions of multivariate test statistics based on the numerical inversion of the associated characteristic functions. These statistics are usually expressed as a linear combination or product of independent random variables with known distributions and characteristic functions. In addition, it is focused on the problem of inversion of multivariate characteristic functions. A numerical algorithm is proposed for the inversion of the bivariate characteristic function, and it is shown how it allows the use of complex probability distributions specified by the characteristic function, including the copula function. The problem of generating random numbers is also discussed when the bivariate distribution is specified by its characteristic function and an algorithm is proposed based on the conditional characteristic function. To illustrate the concept and application of these algorithms, a version of the bivariate logistic distribution specified is used by its characteristic function. The proposed method has been implemented in MATLAB's Characteristic Functions Toolbox (CharFunTool). The approach offers valuable insights into the challenges of inverting multivariate characteristic functions. It provides a promising approach for accurately and efficiently computing the exact statistical distributions in multivariate statistical analysis.

E0421: **Functional ANOVA based on Fourier transform of distribution**
Presenter: **Daniel Hlubinka**, Univerzita Karlova, Czech Republic
Co-authors: Zdenek Hlavka

Functional two-sample tests based on empirical characteristic functionals are studied. A Cramer-von Mises type test statistic is considered with integration over a preselected family of probability measures, say Q, leading to a computationally feasible and powerful test statistic. Small sample properties of the resulting two- and k-sample functional tests are investigated in a simulation study. In particular, the resulting tests are compared to previously proposed tests of equality of mean functions and covariance operators, and it is shown that a proper choice of the probability measure Q gives very good power in detecting shift and scale alternatives.

E0293: **Instabilities in time-dependent implied volatility functional profiles**
Presenter: **Matus Maciak**, Charles University, Czech Republic

A unique methodological approach is proposed to recognize, detect, and estimate a specific type of stochastically relevant (multiple significant) changepoints within a sequence of time-dependent functional profiles, the options' implied volatility (IV) smiles in particular. The main focus is on instabilities caused by various exogenous market effects (induced not by the market itself but rather by some human-made interactions). A robust multivariate semi-parametric estimation is employed to postulate an underlying model that complies with the financial theory on arbitrage-free markets, and a consistent statistical test for detecting significant changepoints is proposed under different theoretical assumptions and practical scenarios. The overall performance is investigated from both the theoretical as well as the empirical perspective. Important applicational issues are addressed using a real data example illustration and finite sample simulations.

EO188 Room 701 TOPICS IN GRAPHICAL MODELING Chair: Sang-Yun Oh

E1167: **Optimal backward-learning approach for Gaussian linear structural equation models**
Presenter: **Gunwoong Park**, Seoul National University, Korea, South

The first optimal backwards-learning approach for Gaussian linear structural equation models (SEMs) is introduced using the best-subset-selection approach. Specifically, the class of optimally identifiable Gaussian linear SEMs is provided. Subsequently, it proves that the proposed algorithm is optimal in terms of the sample complexity. Various simulations verify the theoretical findings and confirm the outstanding performance of the algorithm.

E0313: **Confidence sets for causal discovery**
Presenter: **Mladen Kolar**, University of Chicago, United States

Causal discovery procedures are popular methods for discovering causal structures across the physical, biological, and social sciences. However, most procedures for causal discovery only output a single estimated causal model or single equivalence class of models. A procedure is proposed for quantifying uncertainty in causal discovery. Specifically, linear structural equation models are considered with non-Gaussian errors and propose a procedure that returns a confidence set of causal orderings not ruled out by the data. It is shown that asymptotically, the true causal ordering will be contained in the returned set with some user-specified probability.

E0774: **High-performance computing for sparse graphical models**
Presenter: **Joong-Ho Won**, Seoul National University, Korea, South

High-performance computing (HPC) means computations that are so large that a single (desktop) computer cannot meet their requirement on storage, main memory, and raw computational speed. Computational issues involved with graphical model selection in ultrahigh-dimensional data are discussed, in which the number of variables is about a million. At this scale, HPC on distributed memory systems is often necessary. Communication-avoiding matrix multiplication plays a central role in this setting, provided that a proper method is employed. The experience with communication-avoiding algorithms in two national supercomputing centres is shared.

E0600: **FROSTY: A high-dimensional scale-free Bayesian network learning method**
Presenter: **Sang-Yun Oh**, University of California, Santa Barbara, United States

A scalable Bayesian network learning algorithm is proposed based on sparse Cholesky decomposition. The approach only requires observational data and user-specified confidence levels as inputs and can estimate networks with thousands of variables. The computational complexity of the proposed method is $O(p^3)$ for a graph with p vertices. Extensive numerical experiments illustrate the usefulness of the method with promising results. In simulations, the initial step in the approach also improves an alternative Bayesian network structure estimation method that uses an undirected graph as an input.

EO039 Room 702 RECENT ADVANCES IN EXPERIMENTAL DESIGN AND ANALYSIS Chair: Qian Xiao

E0828: Construction of orthogonal-maxpro latin hypercube designs

Presenter: **Yaping Wang**, East China Normal University, China

Co-authors: Qian Xiao, Sixu Liu

Orthogonal Latin hypercube designs (LHDs) and maximum projection (MaxPro) LHDs are widely used in computer experiments. They are efficient for estimating the trend and Gaussian process (GP) parts of the universal Kriging (i.e. GP) model, respectively, especially when only some of the factors are active. However, the orthogonality is found, and the MaxPro criteria often do not agree with each other. Thus, a new class of optimal designs, orthogonal-MaxPro LHDs, is proposed, optimizing a well-defined multi-objective criterion combining the correlation and the MaxPro metrics. An efficient paralleled algorithm is developed via level permutations and expansions, whose efficiency is guaranteed by theories. Numerical results are presented to show that the construction is fast and the obtained designs are attractive, especially for large computer experiments.

E0923: Maximum one-factor-at-a-time designs for screening in computer experiments

Presenter: **Qian Xiao**, University of Georgia, United States

Identifying important factors from many potentially important factors of a highly nonlinear and computationally expensive black box model is a complex problem. Morris screening and Sobol design are two commonly used model-free methods. A connection between these two seemingly different methods in terms of their underlying experimental design structure and further exploiting this connection is established to develop an improved design for screening called Maximum One-Factor-At-A-Time (MOFAT) design. Efficient methods are also developed for constructing MOFAT designs with a large number of factors. Several examples are presented to demonstrate the advantages of MOFAT designs compared to Morris screening and Sobol design methods.

E1129: Uncertainty quantification of optimal decision in curing process simulation

Presenter: **Qiong Zhang**, Clemson University, United States

The cure processing of composite structures primarily suffers from residual stress inducement. Experimental and numerical studies have shown that an optimum cure cycle is critical to reducing residual stress deformation. However, the relationship between residual stresses and the cure cycle is a black box function involving different sources of uncertainties. A statistical approach is developed to quantify the uncertainty of optimal decisions based on data generated from a curing process simulation and propose experimental design strategies to reduce decision uncertainty.

E1192: Construction of space-filling Latin hypercube designs with flexible run sizes

Presenter: **Sixu Liu**, Yanqi Lake Beijing Institute of Mathematical Sciences and Applications, China

Co-authors: Yaping Wang, Qian Xiao

The purpose is to study using Williams transformed good lattice point (GLP) sets to construct space-filling Latin hypercube designs (LHDs) of size $n\phi(n)$, where $\phi(n)$ is the Euler function and $n = pq$ for distinct primes p and q. Williams transformed GLP sets have recently been shown as a powerful tool for constructing maximin L1-distance LHDs. However, the existing theoretical results only cover the cases of $n = p$ and $n = 2p$. The optimality results for more general sizes of n=pq are shown, where p and q can be any two distinct prime numbers. A simple representation of such GLP designs is derived and used to prove their asymptotic optimality under the maximin L1-distance criterion. A method is also proposed to construct more space-filling LHDs with flexible sizes based on Williams-transformed GLP designs.

EO236 Room 703 STATISTICAL NETWORK ANALYSIS I Chair: Keith Levin

E0777: On network modularity statistics in connectomics and schizophrenia

Presenter: **Joshua Cape**, University of Wisconsin, Madison, United States

Modularity-based methods for structure and community discovery remain popular in the network neuroscience literature and enjoy a history of yielding meaningful neurobiological findings. All the while, the full potential of these methods remains limited in part by an absence of uncertainty quantification guarantees for use in downstream statistical inference. This direction is pursued by revisiting the classical notion of modularity maximization in the analysis of adjacency and correlation matrices. Considering certain latent space network models wherein high-dimensional matrix spectral properties can be precisely analyzed are proposed and argued. Further for the potential usefulness of several new, non-classical modularity-type network statistics. The findings are applied to an analysis of dMRI and fMRI data in the study of schizophrenia. This is based on joint work with Anirban Mitra (Statistics, University of Pittsburgh) and Konasale Prasad (Psychiatry, University of Pittsburgh).

E0893: Conformal prediction for network regression

Presenter: **Robert Lunde**, Washington University in St Louis, United States

Co-authors: Liza Levina, Ji Zhu

A significant problem in network analysis is predicting a node attribute using nodal covariates and summary statistics computed from the network, such as graph embeddings or local subgraph counts. While standard regression methods may be used for prediction, statistical inference is complicated because the nodal summary statistics often exhibit a nonstandard dependence structure. Under a mild joint exchangeability assumption, conformal prediction methods that are finite-sample valid for a wide range of network covariates are shown. A form of asymptotic conditional validity is also proved that is achievable using standard nonparametric regression methods.

E0899: Geometric inference via graph Laplacians

Presenter: **Dena Asta**, The Ohio State University, United States

Networks can be encoded in terms of their graph Laplacians, matrices that intuitively describe the information flow on a network. When those networks are generated from a geometric space X in a certain sense, an intuitive assumption for social networks, then those graph Laplacians are directly approximating some vital information on the space X. Some recent work will be described in inferring the complete geometric structure of X from graph Laplacians under some mild smoothness assumptions on X - an estimator for intrinsic distances in X between the sample points. The key idea underlying this sort of geometric inference is to regard graph Laplacians not merely as linear operators but as linear operators satisfying a specific approximate "product rule" for second derivatives.

E0908: Perturbation analysis of randomized SVD and its applications to high-dimensional statistics

Presenter: **Minh Tang**, North Carolina State University, United States

Co-authors: Yichi Zhang

Randomized singular value decomposition (RSVD) is a class of computationally efficient algorithms for computing the truncated SVD of large data matrices. Given an $n \times n$ symmetric matrix M, the prototypical RSVD algorithm outputs an approximation of the k leading singular vectors of M by computing the SVD of $M^g G$, where g is an integer, and G is a random Gaussian sketching matrix. The statistical properties of RSVD are discussed under a general signal-plus-noise framework, i.e., the observed matrix $\hat{M}$ is assumed to be an additive perturbation of some true but unknown signal matrix M. Upper bounds are first derived for the spectral and $2- > \infty$ norms between the approximate singular vectors of $\hat{M}$ and the true singular vectors of the signal matrix M. These upper bounds depend on the signal-to-noise ratio (SNR) and the number of power iterations g. A phase transition phenomenon is observed in which a smaller SNR requires larger values of g to guarantee convergence of the spectral and $2- > \infty$ distances. Finally, normal approximations are presented for the row-wise fluctuations of the approximate singular vectors and the entrywise fluctuations of the approximate matrix. The theoretical results are illustrated by deriving nearly-optimal performance guarantees for RSVD when applied to three statistical inference problems: community detection, matrix completion, and principal component analysis with missing data.

EO162 Room 704 MODERN DEVELOPMENT IN LONGITUDINAL/SURVIVAL DATA ANALYSES Chair: Hyunkeun Cho

E0468: **Estimation of heterogeneous treatment effect using random forests for competing risks data**
Presenter: **Youngjoo Cho**, Konkuk University, Korea, South
Co-authors: Jaeseong Park

The estimation of heterogeneous treatment effects for uncensored data has been studied extensively. However, efforts to develop methods of estimating heterogeneous treatment effect using machine learning has begun comparatively recently. A novel approach is proposed to estimating heterogeneous treatment effects with respect to cumulative incidence curves in the competing risks data using random forests. The proposed methods employ orthogonal estimating equations and augmented functions based on semiparametric efficiency theory. Simulation studies show the utility of this approach.

E0308: **Statistical approach to handling measurement issues in self-reported data that are longitudinally collected**
Presenter: **MinJae Lee**, University of Texas Southwestern, United States

In the era of precision medicine, researchers are increasingly sensitive to the heterogeneity among at-risk individuals. Evaluating the association between disease progressions and the longitudinal pattern of pharmacological therapy has become more important. However, in many longitudinal studies, self-reported medication usage data collected at patients' follow-up visits could be missing and/or inaccurate/untenable information. These patterns may also dramatically differ between individuals and thus complicate determining the trajectory of medication use and its complete effects on patients. Although traditional existing methods can deal with specific types of missing/incomplete data, inappropriate handling of this complex issue can lead to misleading findings, especially when it depends upon multiple sources of variation over time. A latent class-based statistical approach under the Bayesian quantile regression framework is proposed that incorporates a cluster of unobserved heterogeneity for handling medication usage data with various measurement issues. Findings from the simulation study indicate that the proposed method performs better than traditional methods under certain data distribution scenarios. Applications of the proposed method are also illustrated to real data obtained from the longitudinal study.

E0410: **Bayesian Conway-Maxwell-Poisson regression for longitudinal count data**
Presenter: **Yeongjin Gwon**, University of Nebraska Medical Center, United States
Co-authors: Jane Meza

Longitudinal count data has been widely collected in biomedical research, public health, and clinical trials. These repeated measurements over time on the same subjects need to account for an appropriate dependency. The Poisson regression model is the first choice to model the expected count of interest. However, this may not be appropriate when data exhibit over-dispersion or under-dispersion. Recently, Conway-Maxwell-Poisson (CMP) distribution has been popularly used as the distribution offers the flexibility to capture a wide range of dispersion in the data. Bayesian CMP regression model proposed accommodating over and under-dispersion in modelling longitudinal count data. Specifically, a regression model with random intercept and the slope is developed to capture subject heterogeneity and estimate covariate effects to be different across subjects. A Bayesian computation is implemented via Hamiltonian MCMC (HMCMC) algorithm for posterior sampling. Bayesian model assessment measures are then computed for model comparison. Simulation studies are conducted to assess the accuracy and effectiveness of the methodology. A well-known example of Epilepsy data demonstrates the usefulness of the proposed methodology.

E0662: **A new perspective on unsupervised learning in longitudinal studies**
Presenter: **Hyunkeun Cho**, University of Iowa, United States
Co-authors: Daren Kuwaye, David-Erick Lafontant

The most common use of unsupervised learning is to cluster data into homogeneous groups. As units are measured multiple times, clustering longitudinal data has become popular. Over the last two decades, various clustering methods have been developed to cluster units based on the similarity of longitudinal profiles. new paradigm in longitudinal clustering is introduced, and its potential applications and implications for other clustering methods are discussed. This novel clustering method provides a unique perspective on longitudinal clustering by considering data heterogeneity over time.

EO034 Room 705 RECENT ADVANCE OF HIGH-DIMENSIONAL STATISTICS Chair: Quefeng Li

E0187: **BELIEF in dependence: Leveraging atomic linearity in data bits for rethinking generalized linear models**
Presenter: **Kai Zhang**, University of North Carolina at Chapel Hill, United States
Co-authors: Benjamin Brown, Xiao-Li Meng

Two linearly uncorrelated binary variables must also be independent because non-linear dependence cannot manifest with only two possible states. This inherent linearity is the atom of dependency constituting any complex form of relationship. Inspired by this observation, a framework called binary expansion linear effect (BELIEF) is developed for assessing and understanding arbitrary relationships with a binary outcome. Models from the BELIEF framework are easily interpretable because they describe the association of binary variables in the language of linear models, yielding convenient theoretical insight and striking parallels with the Gaussian world. In particular, an algebraic structure on the predictors with nonzero slopes governs conditional independence properties. With BELIEF, one may study generalized linear models (GLM) through transparent linear models, providing insight into how modelling is affected by choice of link. For example, setting a GLM interaction coefficient to zero does not necessarily lead to the kind of no-interaction model assumption understood under their linear model counterparts. These phenomena are explored, and a host of related theoretical results is provided. Preliminary empirical demonstration and verification of some theoretical results are also provided.

E0215: **Network detection through odds ratio model**
Presenter: **Jinsong Chen**, University of Nevada Reno, United States

A unified modelling approach to network detection through the semi-parametric odds ratio model is proposed. The proposed model is flexible in handling discrete and continuous data, invariant to biased sampling designs, and avoids model incompatibility. A neighbourhood selection approach is proposed and is shown to be sign-consistent under a version of the irrepresentable condition. A coordinate descent algorithm is proposed to solve the computation problem. Simulations demonstrate that the proposed approach has good performance in comparison to the Gaussian modelling approach. The proposed approach is applied to detect the gene-expression network in breast invasive carcinoma.

E0443: **Inference for high-dimensional linear models with locally stationary error processes**

Presenter: **Xiao Guo**, University of Science and Technology of China, China

Linear regression models with stationary errors are well studied, but the non-stationary assumption is more realistic in practice. An estimation and inference procedure for high-dimensional linear regression models with locally stationary error processes is developed. Combined with a proper estimator for the autocovariance matrix of the non-stationary error, the declassified lasso estimator is adopted for the statistical inference of the regression coefficients under the fixed design setting. The consistency and asymptotic normality of the declassified estimators is established under certain regularity conditions. Element-wise confidence intervals for regression coefficients are constructed. The finite sample performance of the method is assessed by simulation and real data analysis.

E0629: **Classification of competing risks under a semiparametric density ratio model with transition of markers**

Presenter: **Huijun Jiang**, University of North Carolina at Chapel Hill, United States

Co-authors: Quefeng Li, Jessica Lin, Feng-Chang Lin

The cause of failure for competing risk data may not always be observable, which imposes additional challenges for estimating the risk of the primary event of interest. In many infectious diseases, episodes of recurrence may arise through a relapse of the initial infection or a new infection. Identifying the true cause of infection is essential to aid in choosing the appropriate treatment. Using time-to-event information, a novel method is presented for classifying the latent cause of failure under a semiparametric density ratio model. The expectation-maximization (EM) algorithm estimates unknown parameters, including the marginal probabilities of the patient-specific causes of failure. In addition, transition likelihoods between covariates at the baseline and at the time of event occurrence are used to provide a better, at least not worse, classification result. The simulation experiments are performed under various scenarios, such as sample size, censoring rate, and approximation methods for estimating the baseline hazard function. The numerical results show that the proposed classifier performs well under all settings. The proposed method is also applied to Cambodia's P. vivax malaria data, classifying recurrent malaria infections as relapse or reinfection.

EO229 Room 708 NEW ADVANCES IN FUNCTIONAL DATA ANALYSIS Chair: Yuhang Xu

E0175: **Testing homogeneity: The trouble with sparse functional data**

Presenter: **Changbo Zhu**, University of Notre Dame, United States

Co-authors: Jane-Ling Wang

Testing the homogeneity between two samples of functional data is an important task. While this is feasible for intensely measured functional data, it is explained why it is challenging for sparsely measured functional data and shows what can be done for such data. In particular, it is shown that testing the marginal homogeneity based on point-wise distributions is feasible under some mild constraints and proposes a new two-sample statistic that works well with both intensively and sparsely measured functional data. The proposed test statistic is formulated upon Energy distance, and the critical value is obtained via the permutation test. The convergence rate of the test statistic to its population version is derived along with the consistency of the associated permutation test. To the best of our knowledge, this is the first paper that provides guaranteed consistency for testing the homogeneity of sparse functional data. The aptness of the method is demonstrated on both synthetic and real data sets.

E0193: **Dynamic modelling for multivariate functional and longitudinal data**

Presenter: **Qixian Zhong**, Xiamen University, China

Co-authors: Siteng Hao, Shu-Chin Lin, Jane-Ling Wang

Dynamic interactions among several stochastic processes are common in many scientific fields. It is crucial to model these interactions to understand the dynamic relationship of the corresponding multivariate processes with their derivatives and to improve predictions. In reality, full observations of the multivariate processes are not feasible as measurements can only be taken at discrete locations or time points and often only sparingly and intermittently in longitudinal studies. This results in multivariate longitudinal data measured at different times for different subjects. A time-dynamic model is proposed to handle multivariate longitudinal data by modelling the derivatives of multivariate processes using the values of these processes. Starting with a concurrent linear model, methods are developed to estimate the regression coefficient functions, which can accommodate irregularly measured longitudinal data that are possibly contaminated with noise. This approach can also be applied to settings where all subjects' observational times are the same. The study establishes the convergence rates of the estimators with phase transitions and further illustrates the proposed model through numerical studies.

E0379: **Learning functional graphical models with additive conditional independence**

Presenter: **Kuang-Yao Lee**, Temple University, United States

Co-authors: Lexin Li, Bing Li, Hongyu Zhao

A nonparametric graphical model is developed for multivariate random functions. Most existing graphical models are restricted by the assumptions of multivariate Gaussian or gaussian copula distributions, which also imply linear relations among the random variables or functions on different nodes. Building our graphical model based on a new statistical object- the functional additive regression operator- relaxed those assumptions. Our method can capture nonlinear relations without requiring distributional assumptions by carrying out regression and neighbourhood selection at the operator level. Moreover, the method is built up using only a one-dimensional kernel, thus avoiding the curse of dimensionality from which a fully nonparametric approach often suffers, and it enables us to work with large-scale networks. Error bounds are derived for the estimated regression operator and establish graph estimation consistency while allowing the number of functions to diverge at the exponential rate of the sample size. The method's efficacy is demonstrated through simulations and an electroencephalography dataset analysis. (This is joint work with Lexin Li (UC Berkeley), Bing Li (Penn State), and Hongyu Zhao (Yale).

E0914: **Approximation, estimation and inferential theory for locally stationary functional time series**

Presenter: **Xiucai Ding**, UC Davis, United States

Some recent results on locally stationary functional time series analysis are reported. First, It is proved that under some mild conditions, every locally stationary functional time series with short-range dependence can be well-approximated by a functional AR process with a diverging number of order which is adaptive to the underlying structures. Second, sieve estimators for the coefficients of the functional AR process, which attains the min-max rates, are provided. Third, inference of these coefficients is conducted by establishing a Gaussian approximation result. Applications include checking the stationarity of the functional time series. A multiplier bootstrap method is proposed for the implementation.

EO016 Room 709 METHODS FOR CAUSAL INFERENCE, PRECISION MEDICINE AND DIMENSION REDUCTION Chair: Zheng Zhang

E0275: **STEEL: Singularity-aware reinforcement learning**

Presenter: **Zhengling Qi**, The George Washington University, United States

Batch reinforcement learning (RL) aims to find an optimal policy in a dynamic environment to maximize the expected total rewards by leveraging pre-collected data. A fundamental challenge behind this task is the distributional mismatch between the batch data-generating process and the distribution induced by target policies. Nearly all existing algorithms rely on the absolutely continuous assumption of the distribution induced by target policies with respect to the data distribution so that the batch data can be used to calibrate target policies via the change of measure. However, the absolute continuity assumption could be violated in practice, especially when the state-action space is large or continuous. A new batch RL algorithm is proposed without requiring absolute continuity in the setting of an infinite-horizon Markov decision process with continuous states and actions. Our algorithm is motivated by a new error analysis on off-policy evaluation, where maximum mean discrepancy, together with distributionally robust optimization, are used to characterize the error of off-policy evaluation caused by the possible singularity and to enable the power of model extrapolation. By leveraging the idea of pessimism and under some mild conditions, a finite-sample regret guarantee is derived for our proposed algorithm without imposing absolute continuity.

E0401: **Casual inference of general treatment effects using neural networks with a diverging number of confounders**
Presenter: **Zheng Zhang**, Renmin University of China, China

Estimating causal effects is a primary goal of behavioural, social, economic and biomedical sciences. Under the unconfoundedness condition, adjustment for confounders requires estimating the nuisance functions relating outcome and/or treatment of confounders. A generalized optimization framework is considered for efficient estimation of general treatment effects using feedforward artificial neural networks (ANNs) when the number of covariates is allowed to increase with the sample size. ANNs estimate the nuisance function, and a new approximation error bound is developed for the ANNs approximators when the nuisance function belongs to a mixed Sobolev space. It is shown that the ANNs can alleviate the curse of dimensionality under this circumstance. The consistency and asymptotic normality of the proposed treatment effects estimators are further established, and a weighted bootstrap procedure for conducting inference is applied. The proposed methods are illustrated via simulation studies and a real data application.

E0481: **Regression adjustment in randomized controlled trials with many covariates**
Presenter: **Yukitoshi Matsushita**, Hitotsubashi University, Japan
Co-authors: Harold Chiang, Taisuke Otsu

The proposed method estimates and predicts average treatment effects in randomized controlled trials when researchers observe potentially many covariates. By employing Neyman's finite population perspective, a bias-corrected regression adjustment estimator using cross-fitting is proposed, and it is shown that the proposed estimator has favourable properties over existing alternatives. For inference, the first and second-order terms are derived in the stochastic component of the regression adjustment estimators, higher-order properties of the existing inference methods are studied, and a bias-corrected version of the HC3 standard error is proposed. Simulation studies show our cross-fitted estimator, combined with the bias-corrected HC3, delivers precise point estimates and robust size controls over a wide range of DGPs. To illustrate, the proposed methods are applied to real datasets on randomized experiments of incentives and services for college achievement following previous researchers.

E0919: **On efficient dimension reduction with respect to the interaction between two response variables**
Presenter: **Wei Luo**, Zhejiang University, China

The theory and the methodologies for dimension reduction concerning the interaction between two response variables are proposed. This is crucial for effective dimension reduction in applications such as missing data analysis and causal inference. The concepts of the locally and the globally efficient dimension reduction subspaces are introduced, which induce reduced predictors that preserve the critical feature for subsequent data analysis. These spaces can be low dimensional when neither of the two individual response variables is equipped with low-dimensional data structures, for which they cannot be recovered by the existing dimension reduction applications in general. Based on the current inverse regression methods, a family of dimension reduction methods is proposed called the dual inverse regression, which consistently estimates the locally efficient dimension reduction subspaces under mild assumptions and consistently estimates the globally efficient dimension reduction subspace when it exists. These methods are also easily implementable. In addition, a sufficient and necessary condition for the existence of the globally efficient dimension reduction subspace that is handy to check is proposed. Simulations studies and a real data example illustrate the usefulness of the proposed dual inverse regression methods.

EP001 Room Poster session II POSTER SESSION II Chair: Cristian Gatu

E0259: **A longitudinal study of the impact of hurricanes in quality of life on women diagnosed with gynecological cancer**
Presenter: **Jan Pena Rivera**, University of Puerto Rico, United States
Co-authors: Istoni Da Luz Santana

As of 2021, cancer was the second leading cause of death in Puerto Rico and the United States Virgin Islands. Gynaecological cancer patients experience numerous stressors, including the physical impact of the disease, treatment side effects, financial difficulties, and many others. On September 2017, Hurricanes Irma and Mara affected Puerto Rico and the U.S. Virgin Islands. Disasters such as these are high-stress events that can lead to negative physical and mental health outcomes for those affected. However, cancer patients are in an especially vulnerable position because of the devastated health infrastructure and limited access to care, among other aspects. Timely and uninterrupted cancer care is essential for cancer survival and quality of life. In order to study the dynamic changes of the effects the disasters (explanatory latent variables) cause in gynaecological cancer patients' quality of life (outcome latent variable), a Bayesian approach using a longitudinal SEM was applied to the data observed before and after the hurricanes Irma and Maria. A Gibbs sampler algorithm was applied to estimate the unknown parameters of the model. The Lv-measure is used as a model comparison statistic. Results obtained are provided to illustrate the Bayesian methodologies.

E0361: **New second-order asymptotic methods for nonlinear models**
Presenter: **Gubhinder Kundhi**, Memorial University of Newfoundland, Canada
Co-authors: Paul Rilstone

New higher-order asymptotic methods for nonlinear models are developed. These include generic methods for deriving stochastic expansions of arbitrary order, methods for evaluating the moments of polynomials of sample averages, a method for deriving the approximate moments of the stochastic expansions, simplified expressions for the first four moments of nonlinear estimators, a third-order approximate Saddlepoint expansion. The techniques are applied to improve inferences with the Two-Stage Least Squares (2SLS) estimator and the weak instruments problem. In a Monte Carlo simulation experiment, higher-order analytical corrections provided by the Saddlepoint approximation are compared to the Edgeworth, the bootstrap and the first-order approximation in finite samples in an i.i.d sampling context.

E0650: **One class classification using Bayesian optimization**
Presenter: **In Young Baek**, Inha University, Korea, South
Co-authors: Seongil Jo, Jae Oh Kim

One Class Classification(OCC) is a technique for detecting abnormal data by creating a decision boundary that defines normal data when there is an imbalance between normal and abnormal data. One Class Support Vector Machine (OC-SVM) and Deep Support Vector Data Description (Deep SVDD) are one of the methodologies of OCC. OCSVM is to find the hyperplane that separates the majority of normal data from the origin in the feature space. Deep SVDD is to find the smallest hypersphere involving the most normal data. By using neural networks, Deep SVDD maps data from the original space to a feature space. OCSVM and Deep SVDD are sensitive to hyperparameters. Methodologies of hyperparameters optimization are Grid Search, Random Search and Bayesian Optimization. The end is to compare the performance between the three methodologies of hyperparameter optimization and show that Bayesian optimization outperforms grid search and random search.

E0715: **Beta regression models: Practical analyses with KNHANES 2013-2015 data and Covid-19 data**
Presenter: **Jeong In Lee**, Inha University, Korea, South
Co-authors: Seong Il Jo, Jae Oh Kim

The continuous response variable with skewness and heteroscedasticity restricted to interval (0,1) violates traditional methods assumptions such as Ordinary Least Squares (OLS). The beta regression model is adequate for these situations. The beta distribution can be parametrized in terms of its mean and precision parameters, and the sub-model for mean can be estimated in beta regression. The variable dispersion beta regression is the extended beta regression model with two submodels for mean and precision. For these beta regression models, estimation can be performed by maximum likelihood or by Bayesian inference. Two practical applications are presented in comparison with traditional linear regression by OLS, beta regression and extended beta regression performed by ML and Bayesian inference. The first practical analysis is applied to Korean National Health and Nutrition Examination Survey (KNHANES) 2013-2015 to investigate the relationship between smoking and coffee intake. The second practical analysis is applied to the Covid-19 data set to examine the association between several county-level characteristics and the cumulative proportions of confirmed cases and deaths in the states of the USA. In these applications, it is illustrated that the beta regression and the extended beta regression are appropriate.

E1013: **Subgroup analysis in the observational studies**

Presenter: **Hoeun Lee**, Konkuk University, Korea, South

Co-authors: Youngjoo Cho

Matching is one of the widely used methods in observational studies to estimate causal effects. In the past decade, the virtual twin's method, the combination of imputation and machine learning method on potential outcomes, has provided many useful results in the estimation of treatment effect and subgroup analysis in the randomized study. Virtual twins are proposed with matching in observational studies for subgroup analysis. The relevance of the proposed method is evaluated by the consistency and accuracy of estimating treatment effects in simulation studies.

E1057: **Tree-structured clustering model using projection pursuit method and their explanation**

Presenter: **Eun-Kyung Lee**, Ewha Womans University, Korea, South

The purpose is to propose a tree-structured clustering model using Projection Pursuit. Projection Pursuit is a dimension reduction methodology that is often used for exploratory analysis of high-dimensional data. Finding the optimal projection vector containing the pattern by the user-defined projection pursuit index is possible. The projection pursuit indices are investigated for unsupervised learning, combined these indices with the tree-structured model for projection pursuit, and proposed the projection pursuit clustering tree. Ten data sets with known class information were used to check the model's performance, and the results for each projection pursuit Index were compared.

E1139: **Improving multiple linear regression with random forest using Mahalanobis distance**

Presenter: **Jaeseong Park**, Korea University, Korea, South

Multiple linear regression is a widely used statistical method for modelling the relationship between a dependent variable and multiple independent variables. Random forest, a popular ensemble learning method, has been shown to be effective in solving complex regression problems. A novel approach is proposed to multiple linear regression using random forest with Mahalanobis distance. Mahalanobis distance is a measure of the distance between a point and a distribution, which takes into account the covariance of the data. By incorporating Mahalanobis distance into the random forest algorithm, for the correlations can be accounted between the independent variables and reduce the influence of outliers. The details of the proposed method are presented and its performance with traditional multiple linear regression and random forest regression is compared.

E1284: **t-distributed stochastic neighborhood embedding of tensor data with two applications**

Presenter: **Soohyun Ahn**, Ajou University, Korea, South

The visualization of high-dimensional data is an essential challenge in numerous fields, and there exists a diverse range of techniques to tackle it. A novel visualization algorithm named matrix t-SNE, addresses the problem of jointly visualizing the rows and columns of matrix-variate data, capturing both row and column features at the same time. The method uses a joint embedding technique that updates both low-dimensional embeddings simultaneously and identifies the nested structure within a particular feature. This is achieved by defining and optimizing a unified loss function that yields a new embedding technique for joint visualization of high-dimensional matrix-variate data in a scatter plot. The proposed algorithm is demonstrated using two real data examples: exergame data and gene expression data.

Thursday 03.08.2023 13:40 - 14:55 Parallel Session N – EcoSta2023

EO211 Room 02 ADVANCES IN STATISTICS Chair: Yang Ni

E0166: **Ordinal regression for non-ordinal data? It's all about parsimony!**
Presenter: **Yang Ni**, Texas AM University, United States

The aim is to introduce a novel regression model for categorical data termed classification with optimal label permutation (COLP). By design, COLP is a more parsimonious classifier than multi-class logistic regression and hence can have better out-of-sample prediction performance. The idea is simple - the class label is ordered so that ordinal regression is applicable. While label order does not have to have any physical meaning, it sometimes does. In addition to classification, interestingly, COLP gives rise to a provably identifiable causal graphical model for categorical data, whereas multi-class logistic regression does not.

E0383: **Accelerating approximate Bayesian computation methods with Gaussian processes**
Presenter: **Shijia Wang**, Nankai University, China

Approximate Bayesian computation (ABC) is a Bayesian inference algorithm class that targets problems with intractable or missing likelihood functions. It approximates the posterior distribution by utilizing simulators to draw synthetic data. However, ABC is computationally intensive for complex models in which simulating synthetic data is very expensive. An early rejection Markov chain Monte Carlo (ejMCMC) sampler is proposed with Gaussian processes to accelerate inference speed. Samples are rejected early in the first stage of the kernel using a discrepancy model, in which the discrepancy between the simulated and observed data is modelled by the Gaussian process (GP). Hence, the synthetic data are generated only if the parameter space is worth exploring. In addition, the proposed method is employed within an ABC sequential Monte Carlo (SMC) sampler. In the numerical experiments, examples of ordinary differential equations, stochastic differential equations, and delay differential equations are used to demonstrate the effectiveness of the proposed algorithm.

E0596: **Exponential-family principal component analysis of two-dimensional functional data with serial correlation**
Presenter: **Kejun He**, Renmin University of China, China
Co-authors: Bohai Zhang, Lan Zhou

Motivated by a study on Arctic sea-ice-extent (SIE) data of binary observations, a novel model is proposed to analyze serially correlated non-Gaussian data observed on a two-dimensional domain that may have an irregular shape. The observed data is assumed to follow a distribution from the exponential family, where the corresponding natural parameter is a dynamic, smooth function of two-dimensional locations. A functional principal component model using bivariate splines defined on triangulations is applied on the natural-parameter surface to characterize the spatial variation of data. Autoregressive (AR) processes model the serial correlation of data observed at consecutive time points on the principal component scores. To estimate the unknown parameters, an EM algorithm is developed with two approaches, using Laplace approximation and variational inference, respectively, on the E-step. Through simulation studies, it is found that the latter is much faster with higher estimation accuracy, especially when the sample size is large. Finally, the proposed model with variational inference EM algorithm is applied to analyze the massive monthly Arctic SIE data.

EO214 Room 03 RECENT DEVELOPMENTS IN COMPLEX DATA ANALYSIS Chair: Xin He

E0972: **Integrative group factor model for variable clustering on temporally dependent data: Optimality and algorithm**
Presenter: **Lyuou Zhang**, Shanghai Univeristy of Finance and Economics, China
Co-authors: Wen Zhou, Haonan Wang

A model-based clustering approach is adopted, in which the population-level clusters are clearly and statistically interpretable to cluster a larger number of variables. The integrative group factor model (iGFM) is proposed, which can handle temporally dependent data and allows for connections across variable clusters. This model introduces two types of latent factors, the common and unique factors, to model cross-cluster connections and within-cluster similarities among variables. The difficulty of clustering variables based on the iGFM in terms of a permutation-invariant clustering risk is quantified, and the minimax signal threshold below is derived, which no algorithms can cluster variables successfully. This threshold is driven by the competition between common and unique factors in the model and does not require a clear separation of clusters to guarantee perfect recovery. Using spectral decomposition and linear search techniques, a fast and minimax-optimal algorithm is developed to cluster variables. An interesting phase transition in the clustering performance is also identified, where the model parameter space is partitioned into three regions corresponding to cases of impossible to cluster perfectly, possible with guarantees on optimality, and possible with no provable guarantees, respectively.

E0974: **Quantile autoregressive conditional heteroscedasticity**
Presenter: **Qianqian Zhu**, Shanghai University of Finance and Economics, China
Co-authors: Songhua Tan, Yao Zheng, Guodong Li

A novel conditional heteroscedastic time series model is proposed by applying the framework of quantile regression processes to the ARCH(∞) form of the GARCH model. This model can provide varying structures for conditional quantiles of the time series across different quantile levels while including the commonly used GARCH model as a special case. The strict stationarity of the model is discussed. For robustness against heavy-tailed distributions, a self-weighted quantile regression (QR) estimator is proposed. While QR performs satisfactorily at intermediate quantile levels, its accuracy deteriorates at high quantile levels due to data scarcity. As a remedy, a self-weighted composite quantile regression (CQR) estimator is further introduced and, based on an approximate GARCH model with a flexible Tukey-lambda distribution for the innovations; the high quantile levels can be extrapolated by borrowing information from intermediate ones. Asymptotic properties for the proposed estimators are established. Simulation experiments are carried out to access the finite sample performance of the proposed methods, and an empirical example is presented to illustrate the usefulness of the new model.

E1109: **Efficient learning of nonparametric directed acyclic graph with statistical guarantee**
Presenter: **Xin He**, Shanghai University of Finance and Economics, China
Co-authors: Yibo Deng, Shaogao Lv

Directed acyclic graph (DAG) models are widely used to represent casual relations among collected nodes. An efficient and consistent method is proposed to learn DAG with a general causal dependence structure, which is in sharp contrast to most existing methods assuming linear dependence of causal relations. The proposed method leverages the concept of a topological layer to facilitate DAG learning. It connects nonparametric DAG learning with kernel ridge regression in a smooth reproducing kernel Hilbert space (RKHS) and learning gradients by showing that the topological layers of a nonparametric DAG can be exactly reconstructed via kernel-based estimation. The parent-child relations can be obtained directly by computing the estimated gradient function. The developed algorithm is computationally efficient in the sense that it attempts to solve a convex optimization problem with an analytic solution, and the gradient functions can be directly computed by using the derivative reproducing property in the smooth RKHS. The asymptotic properties of the proposed method are established in terms of exact DAG recovery without requiring any explicit model specification. Various simulated and real-life examples also support its superior performance.

EO309 Room Virtual R01 RECENT DEVELOPMENTS IN HIGH-DIMENSIONAL CHANGE POINT ANALYSIS Chair: Hyeyoung Maeng

E0435: **High-dimensional dynamic pricing under non-stationarity: Learning and earning with change-point detection**
Presenter: **Zifeng Zhao**, University of Notre Dame, United States

A high-dimensional dynamic pricing problem is considered under non-stationarity, where a firm sells products to T sequentially arriving consumers that behave according to an unknown demand model with potential changes. The demand model is a high-dimensional generalized linear model (GLM), allowing for a feature vector that encodes products and consumer information. To achieve optimal revenue (i.e., least regret), the firm needs to learn and exploit the unknown GLMs while monitoring for potential change points. First, a novel penalized likelihood-based online change-point detection algorithm is designed for high-dimensional GLMs, which is the first algorithm that achieves the optimal minimax localization error rate for high-dimensional GLMs. A change-point detection-assisted dynamic pricing (CPDP) policy is further proposed. It achieves a near-optimal regret of order $O(s\sqrt{\Upsilon_T T}\log(Td))$, where s is the sparsity level, and Υ_T is the number of change points. This regret is accompanied by a minimax lower bound, demonstrating the optimality of CPDP. In particular, the optimality concerning Υ_T is seen for the first time in the dynamic pricing literature and is achieved via a novel accelerated exploration mechanism. Extensive simulation experiments and a real data application on online lending illustrate the efficiency of the proposed policy and the importance and practical value of handling non-stationarity in dynamic pricing.

E0683: **Robust high-dimensional change point detection under heavy tails**
Presenter: **Yudong Chen**, London School of Economics and Political Science, United Kingdom
Co-authors: Mengchu Li, Tengyao Wang, Yi Yu

The mean change point detection problem for heavy-tailed high-dimensional data is studied. Firstly, it is shown that when each component of the error vector follows an independent sub-Weibull distribution, a CUSUM-type statistic achieves the minimax testing rate in almost all sparsity regimes. Secondly, when the error distributions have polynomially decaying tails admitting bounded αth moment for some $\alpha \geq 4$, a median-of-means-type statistic that achieves a near-optimal testing rate in both the dense and the sparse regime is introduced. A "black-box" robust sparse mean estimator is combined with the median-of-means-type statistic to achieve optimality in the sparse regime. Although such an estimator is usually computationally inefficient for its original purpose of mean estimation, the combined approach for change point detection is polynomial time. Lastly, the even more challenging case is investigated when $2 \leq \alpha < 4$ unveil a new phenomenon that the (minimax) testing rate has no sparse regime, i.e. sparse testing changes is information-theoretically as hard as testing dense changes. It is shown that the dependence of the testing rate on the data dimension exhibits a phase transition at $\alpha = 4$.

E1111: **Random forests for change point detection**
Presenter: **Malte Londschien**, ETH Zurich, Switzerland
Co-authors: Peter Buehlmann, Solt Kovacs

Changeforest, a novel nonparametric multivariate change point detection method, is introduced. Change point detection considers the localization of abrupt distributional changes in time series. This has bioinformatics, neuroscience, biochemistry, climatology, and finance applications. The power of modern nonparametric classifiers like random forests is leveraged by reframing the change point detection problem as a supervised learning task. A log-likelihood ratio that uses random forests' class probability predictions is constructed to compare change point configurations and pair this with a computationally feasible search method. It is proved that Changeforest consistently locates change points in single change point scenarios. In a comprehensive simulation study, changeforest achieves improved empirical performance compared to existing methods.

EO013 Room 102 RECENT DEVELOPMENTS IN BAYESIAN METHODS AND HIGH-DIMENSIONAL STATISTICS Chair: Shouzheng Chen

E1209: **Non-segmental Bayesian detection of multiple change-points**
Presenter: **Chong Zhong**, The Hong Kong Polytechnic University, Hong Kong
Co-authors: Zhihua Ma, Xu Zhang, Catherine Liu

Detection of multiple change points has long been important and active in capturing abrupt change signals within various structures of data streams under wide applications. The common spirit of existing literature is segment-wise in the sense that segment parameters of the local signal on each segment are studied segment-wisely. In contrast, a general and original non-segmental approach is proposed. The pure jump process as a global infinite-dimensional parameter is treated and modelled by an atomic representation where random atoms are associated with random heights. Under the atomic representation, the change-point detection is transferred to discriminating the non-zero outliers from the posterior estimates of the jump sizes at all data points. A class of dynamic discrete spike-and-slab shrinkage priors for the random heights in the global parameter is constructed so that the posterior contraction of the jump sizes attains the optimal minimax rate. An empirical 3-sigma criterion is employed to discriminate non-zero jump sizes, resulting in an asymptotically zero false negative rate. In numerical studies, the approach outperforms existing methods in detecting scale shifts and is competent in detecting mean shifts and structural changes under linear regression or auto-regression settings.

E1204: **Estimation of Tucker tensor factor models for high-dimensional higher-order tensor observations**
Presenter: **Xu Zhang**, The Hong Kong Polytechnic University, Hong Kong

Higher-order tensor data prevail in a wide range of fields, including finance and economics, high-resolution videos, multimodality imaging, engineering such as signal processing, and elsewhere. Tucker decomposition may be the most general low-rank approximation method among versatile decompositions of higher-order tensors owing to its strong compression ability, whilst statistical properties of the induced Tucker tensor factor model (TuTFaM) remains a big challenge and yet critical before it provides justification for applications in machine learning and beyond. Existing theoretical developments mainly focus on the field of time series with the assumption of strong auto-correlation among temporally ordered observations, which is ineffective for weakly dependent and independent tensor observations. Under quite mild assumptions, this article kicks off the participation of raw weakly correlated tensor observations within the TuTFaM setting. It proposes two sets of PCA-based estimation procedures, moPCA and its refinement IPmoPCA, the latter of which is enhanced in the rate of convergence. Their asymptotic behaviours, which can reduce to those in low-order tensor factor models in the existing literature, are developed. The proposed approaches outperform existing auto-covariance-based methods for tensor time series in terms of estimation and tensor reconstruction effects in both simulation experiments and two real data examples.

E1297: **Nonparametric transformation models for doubly censored survival data: A Bayesian approach**
Presenter: **Shouzheng Chen**, The Hong Kong Polytechnic University, China
Co-authors: Chong Zhong, Xu Zhang

Doubly censored data are frequently encountered in pharmacological and epidemiological studies, where the failure time can only be observed within a certain range and is otherwise either left or right-censored, and some analysis and inference procedures have been established. Predictions of survival times are made for doubly censored data under a nonparametric transformation model where both the monotone transformation function and the model error are unspecified. The nonparametric transformation model is robust to model misspecification, leading to stable predictions in various practical settings. Bayesian inference is facilitated without identifying the model by constructing weakly informative nonparametric priors for the infinite-dimensional parameters. Considering the left and right censoring times for real-life doubly censored data are commonly fixed, for the transformation function, a pseudo-quantile I-splines prior is proposed, which places interior knots of I-spline functions at average quantiles of synthetic doubly censored data. Such a prior characterizes the major body of the transformation function well and outperforms commonly used I-splines priors with equally spaced knots. Comprehensive simulations and an application to an AIDS clinical trial demonstrate that the proposed method outperforms existing approaches.

EO058 Room 201 CAUSAL INFERENCE Chair: Yen-Tsung Huang

E0738: **G-estimation with invalid instrumental variables**
Presenter: **BaoLuo Sun**, National University of Singapore, Singapore
Co-authors: Zhonghua Liu, Eric Tchetgen Tchetgen

The instrumental variable method is widely used in the health and social sciences to identify and estimate causal effects in the presence of potentially unmeasured confounding. Multiple instruments are routinely used to improve efficiency, leading to concerns about bias due to possible violation of the instrumental variable assumptions. To address this concern, a new class of g-estimators that are guaranteed is introduced to remain consistent and asymptotically normal for the causal effect of interest provided that a set of at least k out of K candidate instruments are valid, for some value of k set by the analyst ex-ante, without necessarily knowing the identities of the valid and invalid instruments.

E1064: **Sparse quantile regression**
Presenter: **Le-Yu Chen**, Institute of Economics, Academia Sinica, Taiwan
Co-authors: Sokbae Lee

Both L0-penalized and L0-constrained quantile regression estimators are considered. For the L0-penalized estimator, an exponential inequality on the tail probability of excess quantile prediction risk is derived, and it is applied to obtain non-asymptotic upper bounds on the mean-square parameter and regression function estimation errors. Analogous results for the L0-constrained estimator are also derived. The resulting rates of convergence are nearly minimax-optimal, and the same as those for L1-penalized and non-convex penalized estimators. Further, the expected Hamming loss for the L0-penalized estimator is characterized. The proposed procedure is implemented via mixed integer linear programming and also a more scalable first-order approximation algorithm. The finite-sample performance of the proposed approach in Monte Carlo experiments and its usefulness in a real data application concerning the conformal prediction of infant birth weights (with n 1000 and up to p>1000) are illustrated. In sum, the L0-based method produces a much sparser estimator than the L1-penalized, and non-convex penalized approaches without compromising precision.

E1119: **Separable effects under semicompeting risks**
Presenter: **Jih-Chang Yu**, Academia Sinica, Taiwan
Co-authors: Yen-Tsung Huang

The separable effect has recently been proposed to study the causal effects under the setting of competing risks. The separable effect approach is extended to the semi-competing risks involving a primary outcome and an intermediate outcome. The exposure into two disjoint components is decomposed: the first component affects the primary outcome directly, i.e., direct effect and the other affects the primary outcome through the intermediate outcome, i.e., indirect effect. Under such effect separation, the identification formula of counterfactual risk derived for semi-competing risks is a function of cause-specific hazards and transition hazards of multistate models. It can be reduced to the formula for competing risks as a special case. Both nonparametric (NP) and semiparametric (SP) methods are proposed to estimate the causal effects and study their asymptotic properties. The model-free NP method is robust but less efficient for confounder adjustment; the model-based SP method flexibly accommodates confounders by treating them as covariates. Comprehensive simulations are conducted to study the performance of the proposed methods. Finally, the proposed methods are applied to characterize how hepatitis C infection affects the incidence of liver cancer through liver cirrhosis.

EO219 Room 203 CHECKING FOR MODEL STRUCTURAL CHANGE IN HIGH-DIMENSIONAL DATA Chair: Tiejun Tong

E0402: **Multiple change point detection in tensors**
Presenter: **Jiaqi Huang**, Beijing Normal University, China
Co-authors: Junhui Wang, Lixing Zhu, Xuehu Zhu

A criterion is proposed for detecting change structures in tensor data. To accommodate tensor structures that may have a structural mode that is not suitable to be summarized in a distance to measure the difference between any two adjacent tensors, a mode-based signal-screening Frobenius distance for the moving sums of slices of tensor data is defined to handle both dense and sparse model structures of the tensors. It can also deal with the case without structural mode as a general distance. Based on the distance, signal statistics are then constructed using the ratios with adaptive-to-change ridge functions. The number of changes and their locations can then be consistently estimated in certain senses, and the confidence intervals of the locations of change points are constructed. The results hold when the size of the tensor and the number of change points diverge at certain rates, respectively. Numerical studies are conducted to examine the finite sample performances of the proposed method. Two real data examples are also analyzed for illustration.

E0835: **Testing the martingale difference hypothesis in high dimension**
Presenter: **Qing Jiang**, Beijing Normal University, China

The aim is to test the martingale difference hypothesis for high-dimensional time series. The test is built on the sum of squares of the element-wise max-norm of the proposed matrix-valued nonlinear dependence measure at different lags. To conduct the inference, the null distribution of the test statistic is approximated by Gaussian approximation and a simulation-based approach to generate critical values is provided. The asymptotic behaviour of the test statistic under the alternative is also studied. This approach is nonparametric as the null hypothesis only assumes the time series concerned is martingale difference without specifying any parametric forms of its conditional moments. As an advantage of Gaussian approximation, the test is robust to the cross-series dependence of unknown magnitude. To the best of our knowledge, this is the first valid test for the martingale difference hypothesis that not only allows for large dimensions but also captures nonlinear serial dependence. The practical usefulness of the test is illustrated via simulation and real data analysis. The test is implemented in a user-friendly R-function.

E1268: **Multiple change point detection for high-dimensional data**
Presenter: **Wenbiao Zhao**, Hong Kong Baptist University, China
Co-authors: Lixing Zhu, Falong Tan

The purpose is to investigate simultaneously detecting multiple change points of the high-dimensional data, with either sparse or dense structure, that the dimension can be of the exponential rate of the sample size. The proposed estimation approach utilizes a signal statistic based on a sequence of signal screening-based local U-statistics. It can avoid both expensive computations that exhaustive search algorithms need and false positives that hypothesis testing-based approaches have to control. The estimation consistency can hold for the locations and number of change points even when the number of change points diverges at a certain rate as the sample size goes to infinity. Further, because of its visualization nature, in practice, plotting the signal statistic can greatly help identify the locations in contrast to existing methods in the literature. Numerical studies are conducted to examine its performance in finite sample scenarios, and a real data example is analyzed for illustration.

EO069 Room 503 RECENT ADVANCES IN BAYESIAN ANALYSIS Chair: Jouchi Nakajima

E0424: **Bayesian analysis of verbal autopsy data using probit model with age- and sex-dependent association between symptoms**
Presenter: **Tsuyoshi Kunihama**, Kwansei Gakuin University, Japan

Verbal autopsy surveys have been used for understanding distributions of deaths by cause, which is fundamental public health information, in low-resource settings without well-organized vital statistics systems. A new Bayesian approach which extracts the information of distributions of causes of death is developed from verbal autopsy data by taking into account its feature that associations between symptoms vary over the age and sex of individuals. Using gold-standard verbal autopsy data from the Population Health Metrics Research Consortium, we assess the performance of the proposed method by comparing it with existing approaches in this literature. Further, we evaluate the importance of predictors based on information-theoretic measures.

E0968: **Gibbs sampler for matrix generalized inverse Gaussian distributions**
Presenter: **Kaoru Irie**, University of Tokyo, Japan
Co-authors: Shonosuke Sugasawa, Yasuyuki Hamura

Sampling from matrix generalized inverse Gaussian (MGIG) distributions is required in Markov Chain Monte Carlo algorithms for a variety of statistical models. However, an efficient sampling scheme for the MGIG distributions has not been fully developed. Here a novel blocked Gibbs sampler is proposed for the MGIG distributions based on the Choleski decomposition. It is shown that the full conditionals of the diagonal and unit lower-triangular entries are univariate generalized inverse Gaussian and multivariate normal distributions, respectively. Several variants of the Metropolis-Hastings algorithm can also be considered for this problem, but the average acceptance rates are mathematically proved that become extremely low in particular scenarios. The computational efficiency of the proposed Gibbs sampler is demonstrated through simulation studies and data analysis.

E0737: **Bayesian estimation of R-vine Copula with Gaussian-mixture GARCH margins**
Presenter: **Nuttanan Wichitaksorn**, Auckland University of Technology, New Zealand
Co-authors: Rewat Khanthaporn

The purpose is to show the Bayesian estimation of multivariate regular vine (R-vine) copula models with the generalized autoregressive conditional heteroskedasticity (GARCH) margins having the Gaussian-mixture distributions. The Bayesian estimation consists of Markov chain Monte Carlo (MCMC) and variational Bayes (VB) with data augmentation. Due to expensive computation, R-vines have been of limited use while this issue is overcome through parallel computation. To illustrate this, thirteen bivariate copula functions are used for an R-vine pair structure with a large number of marginal distributions where the exponential-type GARCH margins are modelled with the intertemporal capital asset pricing specification through the mixture of Gaussian and generalized Pareto distributions. Results from a simulation study indicate that the proposed models and methods outperform the competing ones. With 100 financial returns in an empirical study, favourable results are still obtained.

EO110 Room 506 RECENT ADVANCES IN HIGH-DIMENSIONAL ECONOMETRICS Chair: Qingliang Fan

E0393: **On the instrumental variable estimation with many weak and invalid instruments**
Presenter: **Qingliang Fan**, The Chinese University of Hong Kong, Hong Kong

The fundamental issue of identification in linear instrumental variable (IV) models with unknown IV validity is discussed. The popular majority and plurality rules are revisited, and the identification conditions, in general, are discussed. With the assumption of the sparsest rule, which is equivalent to the plurality rule but becomes operational in computation algorithms, the advantages of non-convex penalized approaches over other IV estimators based on two-step selections are investigated and proved in terms of selection consistency and accommodation for individually weak IVs. Furthermore, a surrogate sparsest penalty is proposed that aligns with the identification condition and provides an oracle sparse structure simultaneously. Desirable theoretical properties are derived for the proposed estimator with weaker IV strength conditions compared to the previous literature. Finite sample properties are demonstrated using simulations, and the selection and estimation method is applied to an empirical study concerning the effect of trade on economic growth.

E0623: **GMM estimation for high-dimensional panel data models**
Presenter: **Tingting Cheng**, Nankai University, China

A class of high dimensional moment restriction panel data models with interactive effects are studied, where factors are unobserved, and factor loadings are nonparametrically unknown smooth functions of individual characteristics variables. The dimension of the parameter vector and the number of moment conditions is allowed to diverge with the sample size. This general framework includes many existing linear and nonlinear panel data models as special cases. A sieve-based generalized method of the moments estimation method is proposed to estimate the unknown parameters, factors and factor loadings. It is shown that all those unknown quantities can be consistently estimated under a set of simple identification conditions. Further, asymptotic distributions of the proposed estimators are established. In addition, tests for over-identification and specification of factor loading functions are proposed, and their large sample properties are established. Moreover, a number of simulation studies are conducted to examine the performance of the proposed estimators and test statistics in finite samples. An empirical example of stock return prediction is studied to demonstrate the usefulness of the proposed framework and corresponding estimation methods and testing procedures.

E0730: **A conditional linear combination test with many weak instruments**
Presenter: **Wenjie Wang**, Nanyang Technological University, Singapore
Co-authors: Yichong Zhang, Dennis Lim

A linear combination of jackknife Anderson-Rubin (AR), jackknife Lagrangian multiplier (LM), and orthogonalized jackknife LM tests are considered for inference in IV regressions with many weak instruments and heteroskedasticity. Following previous work, the weights in the linear combination are chosen based on a decision-theoretic rule that is adaptive to the identification strength. Under both weak and strong identifications, the proposed test controls the asymptotic size and is admissible among a certain class of tests. Under strong identification, the linear combination test has optimal power against local alternatives among the class of invariant or unbiased tests, which are constructed based on jackknife AR and LM tests. Simulations and an empirical application confirm the good power properties of our test.

EO043 Room 603 ADVANCES IN LARGE-SCALE INFERENCE Chair: Bradley Rava

E0697: **Large scale partial correlation screening**
Presenter: **Peter Radchenko**, University of Sydney, Australia

Identifying multivariate dependencies in high-dimensional data is an important problem in large-scale inference. This problem has motivated recent advances in mining (partial) correlations, focusing on the challenging ultra-high dimensional setting where the sample size n is fixed while the number of features p grows without bounds. A novel principled framework for partial correlation screening with error control will be discussed, leveraging the connection between partial correlations and regression coefficients. Inferential properties of the proposed approach will be established when n is fixed, and p grows to infinity. The theory and methods will be validated on simulated and real data.

E0771: **Unrestricted hypothesis testing**
Presenter: **Bradley Rava**, University of Sydney, Australia

A novel multiple-hypothesis testing approach is presented that improves upon existing state-of-the-art benchmarks by directly utilising p-values to perform calibration to control the local FDR. The methodology has broad applications across multiple fields and works in a variety of settings. The key insights and findings of the research will be discussed.

E0855: **Value-at-Risk forecasts under misspecified conditional models**
Presenter: **Ye Chen**, The University of Sydney, Australia

Co-authors: Bradley Rava, Nam Ho-Nguyen

GARCH-type models are commonly used to capture the dynamics of conditional volatility in financial time series. Vast empirical evidence suggests that the conditional distributions of financial returns tend to be heavy-tailed and asymmetric. To avoid the challenging task of finding the correct parametric family of innovation distributions, Gaussian quasi-maximum likelihood estimators are frequently used to obtain consistent parameter estimates. However, such misspecified conditional models will lead to unsatisfactory Value-at-Risk (VaR) forecasts. A method is provided for obtaining adequate VaR forecasts under models with misspecified innovation distributions.

EO042 Room 604 CRYPTOCURRENCY, RENEWABLE ENERGY AND EFFICIENCY Chair: Artem Prokhorov

E1147: **Econometric modelling of cryptocurrency prices**
Presenter: **Takamitsu Kurita**, Kyoto Sangyo University, Japan
Co-authors: Jennifer L Castle

The rapid expansion of the global cryptocurrency market raises the question of whether there are stable relationships between the prices of representative cryptocurrencies and economic indicators capturing expectations of future monetary policy. Multivariate time series analysis reveals a single but significant cointegrating relationship between several cryptocurrencies and an interest rate spread. This evidence reveals direct implications for timplementingmonetary policy, allowing for the growing influence of digital assets. A policy simulation study using an empirical cointegrated system is conducted to shed light on the controllability of one of the modelled cryptocurrency prices.

E1241: **Iterative distributed multinomial logistic regression**
Presenter: **Xuetao Shi**, the University of Sydney, Australia
Co-authors: Yanqin Fan, Yigit Okar

An iterative estimator for the multinomial logistic regression model is introduced that is both asymptotically efficient and fast to compute even when the number of choices is large. In many economic applications, such as text analysis and spatial choice models, the number of discrete choices can be large. Solving for the maximum likelihood estimator via traditional optimization algorithms, such as Newton-Raphson, is infeasible because the number of arguments in the log-likelihood function is enormous. This problem is tackled by proposing an iterative estimator that optimizes the two parts of the log-likelihood function in turn. The proposed estimator allows for distributed computing, which substantially reduces the computational time. It is shown that the estimator is consistent and has the same asymptotic distribution as the maximum likelihood estimator. Via an extensive simulation study, it is shown that the iterative estimator has good finite sample performance and is extremely fast to compute.

E1237: **Dependence in models of production: New estimators and techniques**
Presenter: **Artem Prokhorov**, University of Sydney, Australia

The core methodological and empirical developments surrounding stochastic frontier models are covered that incorporate various new forms of dependence. Such models apply naturally to panels where cross-sectional observations on firm productivity correlate over time, but also in situations where various components of the error structure correlate with each other and with input variables. Ignoring such dependence patterns is known to lead to severe biases in the estimates of production functions and to incorrect inference. Some of the new approaches are based on techniques from statistical machine learning.

EO142 Room 605 STATISTICAL INFERENCE FOR NONSTATIONARY DATA STRUCTURES Chair: Claudio Durastanti

E0220: **Change point inference in high-dimensional regression models under temporal dependence**
Presenter: **Daren Wang**, University of Notre Dame, United States

Detecting when the underlying distribution changes for the observed time series is a fundamental problem in a broad spectrum of applications. The limiting distributions of change point estimators in the high-dimensional linear regression time series context are considered. At unknown time points, called change points, the regression coefficients change, with the jump sizes measured in l2-norm. Limiting distributions of the change point estimators will be discussed in the regimes where the minimal jump size vanishes and remains constant. The covariate and noise sequences are allowed to be 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 the functional dependence framework, which facilitates the practical implementation of the derived limiting distributions. Extensive numerical results are provided to support the theoretical findings.

E0578: **Spherical autoregressive multiple change-point detection**
Presenter: **Federica Spoto**, Sapienza University of Rome, Italy
Co-authors: Alessia Caponera, Pierpaolo Brutti

Spatio-temporal processes arise very naturally in a number of different applied fields, like Cosmology, Astrophysics, Geophysics, Climate and Atmospheric Science. In most areas, detecting structural breaks or regime shifts in the data stream is key. To this end, the method aims at generalizing the recently introduced SPHAR(p) process by allowing for temporal changes in its functional parameters and variability structure. The approach, which intrinsically integrates the spatial and temporal dimensions, could give multiscale insights into the global and local behaviour of changes, and its performance will be tested on a real dataset of global surface temperature anomalies.

E0761: **Graphical models for nonstationary time series**
Presenter: **Sumanta Basu**, Cornell University, United States
Co-authors: Suhasini Subbarao

NonStGM is proposed, a general nonparametric graphical modelling framework for studying dynamic associations among nonstationary multivariate time series components. It builds on the framework of Gaussian Graphical Models (GGM) and stationary time series Graphical models (StGM). It complements existing works on parametric graphical models based on change point vector autoregressions (VAR). Analogous to StGM, the proposed framework captures conditional noncorrelations (both intertemporal and contemporaneous) in the form of an undirected graph. In addition, the new notion of conditional nonstationarity/stationarity is introduced and incorporated within the graph. This can be used to search for small subnetworks that serve as the source of nonstationarity in a large system. Conditional noncorrelation and stationarity between and within the multivariate time series components to zero and Toeplitz embeddings of an infinite-dimensional inverse covariance operator are explicitly connected. In the Fourier domain, conditional stationarity and noncorrelation relationships in the inverse covariance operator are encoded with a specific sparsity structure of its integral kernel operator. It is shown that these sparsity patterns can be recovered from finite-length time series by node-wise regression of discrete Fourier Transforms (DFT) across different Fourier frequencies. The feasibility of learning the NonStGM structure from data using simulation studies is demonstrated.

EO084 Room 606 ADVANCES IN BUSINESS ANALYTICS Chair: Amanda Chu

E0960: **Graphical copula GARCH modelling with dynamic conditional dependence**
Presenter: **Shun Hin Chan**, The Hong Kong University of Science and Technology, Hong Kong
Co-authors: Amanda Chu, Mike So

The aim is to develop a graphical copula GARCH model for volatility modelling. To allow high-dimensional modelling for large portfolios, the complexity of the modelling is greatly reduced by introducing conditional independence among stocks given the market risk factors, such as the

S&P500 index in the United States. The market risk factors are modeled using a directed acyclic graph (DAG) model with a pairwise-copula construction to allow flexible distributional modelling. Using the DAG model gives a topological order to the market risk factors, which can be regarded as a list of directions of the flow of information or disturbance. The conditional distributions among stock returns are also modelled through pairwise-copula constructions for flexibility. Dynamic conditional dependence structures are adapted to allow the parameters in the copulas to be time-varying such that the tail dependence can dynamically be modelled between any two stocks. Three-stage estimation is used for estimating parameters in the marginal distributions, the copulas of the DAG of the market risk factors, and the copulas of the stocks. Bayesian inference is used to learn the structure of the DAG. The simulation study shows that these estimation procedures can be used to recover the parameters and the DAG accurately. With Bayesian inference, the structure of the market risk factors can be allowed to be random, and model averaging can be done to obtain robust volatility predictions.

E1028: **A semi-parametric multidimensional and longitudinal item response model with mixed data type**
Presenter: **Thomas Chan**, Hong Kong University of Science and Technology, Hong Kong
Co-authors: Mike So, Amanda Chu

A semi-parametric multidimensional and longitudinal item response model with mixed data types is studied. The semi-parametric model consists of a non-parametric item characteristic curve and a parametric performance measurement. The combination of these two parts takes the balance between model flexibility and interpretability. The multidimensional model measures multiple features observed from the data. The longitudinal model captures the changes of respondents throughout the study. When the dataset is large, multiple types of data are often involved. The estimation process is also computationally inefficient. Variational Bayes searches for the best distribution from a pre-assigned family to approximate the complex posterior distribution. The aim is at an efficient solution while maintaining accuracy via variational Bayes. An application of this model is considered, and the performance on estimation with variational Bayes is demonstrated.

E1134: **The development of an automatic speech analytics program to detect the level of stress burden**
Presenter: **Jacky Ngai Lam Chan**, The Hong Kong University of Science and Technology, Hong Kong
Co-authors: Amanda Chu, Mike So

The stress burden generated from family caregiving makes caregivers particularly prone to developing psychosocial health issues; however, with early diagnosis and intervention, disease progression and long-term disability can be prevented. An automatic speech analytics program (ASAP) was developed for the detection of psychosocial health issues based on clients' speech. One hundred Cantonese-speaking family caregivers were recruited. The results suggest that the ASAP can identify family caregivers with low or high-stress burden levels with an accuracy rate of 72%. The findings indicate that digital health technology can be used to assist in psychosocial health assessment. While the conventional method requires rigorous assessments by specialists with multiple rounds of questioning, the ASAP can provide a cost-effective and immediate initial assessment to identify high levels of stress among family caregivers so they can be referred to social workers and healthcare professionals for further assessments and treatments.

EO232 Room 701 RECENT DEVELOPMENTS IN ECONOMETRIC THEORY Chair: Cy Sin

E0560: **Nonparametric and semiparametric estimation of upward rank mobility curves**
Presenter: **Tsung-Chih Lai**, National Chung Cheng University, Taiwan
Co-authors: Jia-Han Shih, Yi-Hau Chen

A novel approach to measuring upward mobility in income ranks across generations that considers the heterogeneity within income classes is presented. Specifically, a previous measure is extended to its continuous form, and a tuning parameter-free nonparametric estimator based on the empirical beta copula is proposed. The estimator is shown to be a particular case of the empirical Bernstein copula-based estimator, with all polynomial degrees equal to the sample size. In addition, a semiparametric distribution regression-based estimator is suggested for conditional mobility with unconditional (or fixed) ranks, which converges weakly to a Gaussian process at the parametric rate and has better finite-sample properties. Applying these methods to the National Longitudinal Survey of Youth data shows strong evidence of stochastic dominance relations in upward rank mobility between blacks and whites in the United States.

E0493: **Testing the impacts on inefficiency in a semiparametric stochastic frontier model**
Presenter: **Jen-Che Liao**, National Chengchi University, Taiwan

The aim is to deal with the significance testing of the effects of exogenous determinants upon the one-sided deviation term of a semiparametric stochastic frontier model. Two nonparametric significance tests for all or a subset of the determinants of inefficiency are proposed. The proposed tests are based on conditional moment restrictions and stochastic processes, with critical values being simulated by means of a multiplier bootstrap procedure. The testing methodology addresses the omitted variable bias that arises naturally in stochastic frontier models when accommodating the determinants of inefficiency and accounts for the estimation effects that appear by using the estimated composite error when constructing the test statistics. The theoretical properties of the proposed tests and the resampling approximations are investigated. The proposed tests are illustrated through simulation experiments and two empirical examples in which the hypotheses of no impacts on inefficiency need to be tested.

E0793: **Inference on three-pass regression filter with high-dimensional target variables**
Presenter: **Shou-Yung Yin**, National Taipei University, Taiwan

The focus is on the high-dimensional target variables using the three-pass regression filter (3PRF) proposed previously. The 3PRF is designed for extracting useful information from an extra data set, which is supposed to be useful for improving the forecast performance of the target variable. Compared to the traditional principal component approach (PCA) based on eigendecomposition, the 3PRF delivers a closed-form solution of the estimated factors and the corresponding estimated coefficients. The proposed approach is robust to the different presumed factor numbers, while the performance of the PCA approach is very sensitive to the number of factors. In the empirical study, the proposed method is used to extract the common components which can be used to predict the fundamentals of the dynamics of house prices in the U.S.

EO097 Room 702 MODERN STATISTICAL METHODS FOR LONGITUDINAL AND IMAGING DATA Chair: Frederic Ferraty

E0347: **Functional concurrent hidden Markov models**
Presenter: **Xinyuan Song**, Chinese University of Hong Kong, Hong Kong

Functional concurrent hidden Markov models are considered. The proposed model consists of two components. One is a transition model for elucidating how potential covariates influence the transition probability from one state to another. The other is a conditional functional linear concurrent regression model for characterizing the state-specific effects of functional covariates. A distribution-free random effect is introduced to the conditional model to describe the dependency of individual functional observations. The soft-thresholding operator and the adaptive group lasso are introduced to simultaneously accommodate the local and global sparsity of the functional coefficients. A Bayesian approach is developed to jointly conduct estimation, variable selection, and the detection of zero-effect regions. This proposed approach incorporates the dependent Dirichlet process with stick-breaking prior for accommodating the unspecified distribution of the random effect and a blocked Gibbs sampler for efficient posterior sampling. Finally, the empirical performance of the proposed method is evaluated through simulation studies, and an application to the analysis of air pollution and meteorological data demonstrates the utility of the methodology.

E0576: **Order selection for heterogeneous semiparametric hidden Markov models**
Presenter: **Yudan Zou**, The Chinese University of Hong Kong, Hong Kong

Hidden Markov models (HMMs), which can characterize dynamic heterogeneity, are useful instruments for analyzing longitudinal data. The order of HMMs, or the number of hidden states, is typically assumed to be known or predetermined by criterion-based techniques in conventional analysis. Considering pairwise comparisons under criterion-based methods become computationally expensive as the model space expands, a few studies have conducted order selection and parameter estimation simultaneously. However, because they only considered homogeneous parametric instances, they cannot account for circumstances in which non-parametric forms or heterogeneity appear. The aim is to propose a Bayesian double-penalized (BDP) procedure for simultaneous order selection and parameter estimation for heterogeneous semi-parametric HMMs. To overcome the difficulties in updating the order, a brand-new Markov chain Monte Carlo algorithm and an effective adjust-bound reversible jump strategy are created. Simulation results reveal that the proposed BDP procedure performs well in estimation and works noticeably better than the standard criterion-based approaches. Application of the suggested method to the Alzheimer's Disease Neuroimaging Initiative research further supports its usefulness.

E0577: **Bayesian quantile scalar on image quantile regression via a nonparametric method**
Presenter: **Chuchu Wang**, The Chinese University of Hong Kong, Hong Kong

The motivation comes from research on Alzheimer's disease and the use of medical imaging data. A quantile scalar-on-image regression model enables a comprehensive study of the relationship between cognitive decline and various clinical covariates and imaging factors. A Bayesian nonparametric model is used here to cope with the complex spatially distributed imaging data. It is assumed that there is a latent Gaussian process to capture the sparse structure of the regression coefficients, and the soft-thresholded operator is used to shrink the coefficient function. The kernel basis functions approximate the latent Gaussian process, which promises an efficient MCMC computation algorithm and a consistent estimation result. The model is represented in a hierarchical form, and a fully Bayesian approach is constructed. A hybrid algorithm combining the Gibbs sampler and Metropolis-Hastings algorithm is adopted to conduct posterior sampling and make the inference. The method's performance is compared with the functional principal component analysis (FPCA) method in simulations, and finally, the proposed method is applied to a study of Alzheimer's disease.

EO235 Room 703 STATISTICAL NETWORK ANALYSIS II Chair: Avanti Athreya

E0631: **Estimating the prevalence of peer effects in network experiments**
Presenter: **David Choi**, Carnegie Mellon University, United States

In randomized experiments with arbitrary and unknown interference, such as social network settings where the treated and control units may affect each other, recent work has proposed hypothesis tests for the null hypothesis of no interference, i.e., that unit is unaffected by the treatment of others. However, without further assumptions, rejection of this null only implies that at least one individual was affected by a treatment other than their own. It is shown that these tests can be inverted with no assumptions on the nature of the interference, producing one-sided interval estimates (or lower bounds) not for the peer effect itself but rather for the number of units affected by the treatment of others. This does not fully identify a peer effect but may be used to show that it exists and estimate whether it is widely prevalent.

E0881: **Estimating network-mediated causal effects via spectral embeddings**
Presenter: **Keith Levin**, University of Wisconsin, United States

The last several years have seen a renewed and concerted effort to incorporate network data into standard regression analysis tools and make network-linked data legible to working scientists. Thus far, this literature has primarily developed tools to infer associative relationships between nodal covariates and network structure. statistical model is augmented for network regression with counterfactual assumptions. Under this model, causal effects can be partitioned into a direct effect not influenced by the network and an indirect effect induced by homophily. The method is a conceptually straightforward integration of latent variable models for networks into the well-known product-of-coefficients mediation estimator. This method is semi-parametric, easy to implement, and highly scalable.

E0906: **Beyond the adjacency matrix: Random line graphs and inference for networks with edge attributes**
Presenter: **Zachary Lubberts**, Johns Hopkins University, United States
Co-authors: Avanti Athreya, Carey Priebe, Youngser Park

Any modern network inference paradigm must incorporate multiple aspects of network structure, including information often encoded in vertices and edges. Methodology for handling vertex attributes has been developed for a number of network models, but comparable techniques for edge-related attributes remain largely unavailable. This gap is addressed in the literature by extending the latent position random graph model to the line graph of a random graph, which is formed by creating a vertex for each edge in the original random graph and connecting each pair of edges incident to a common vertex in the original graph. Concentration inequalities are proved for the spectrum of a line graph and then establish that although naive spectral decompositions can fail to extract the necessary signal for edge clustering, there exist signal-preserving singular subspaces of the line graph that can be recovered through a carefully-chosen projection. Moreover, edge latent positions can be consistently estimated in a random line graph, even though such graphs are of random size, typically have a high rank, and possess no spectral gap. The results also demonstrate that the line graph of a stochastic block model exhibits underlying block structure, and the methods are synthesized and tested in simulations for cluster recovery and edge covariate inference in stochastic block model graphs.

EO213 Room 704 RECENT ADVANCES IN FINANCIAL BIG DATA ANALYSIS Chair: Minseok Shin

E1007: **Robust high-dimensional time-varying coefficient estimation**
Presenter: **Minseok Shin**, KAIST, Korea, South
Co-authors: Donggyu Kim

A novel high-dimensional coefficient estimation procedure based on high-frequency data is developed. Unlike usual high-dimensional regression procedures such as LASSO, the heavy-tailedness of high-frequency observations as well as time variations of coefficient processes, are additionally handled. Specifically, Huber loss and truncation schemes are employed to handle heavy-tailed observations, while l1-regularization is adopted to overcome the curse of dimensionality under a sparse coefficient structure. To account for the time-varying coefficient, local high-dimensional coefficients are estimated, which are biased estimators due to the l1-regularization. Thus, when estimating integrated coefficients, a debiasing scheme is proposed to enjoy the law of large number property and a thresholding scheme is employed to accommodate the sparsity of the coefficients further. This Robust is called thrEsholding Debiased LASSO (RED-LASSO) estimator. It is shown that the RED-LASSO estimator can achieve a near-optimal convergence rate with only a finite b-th moment for any $b > 2$. In the empirical study, the RED-LASSO procedure is applied to the high-dimensional integrated coefficient estimation using high-frequency trading data.

E1017: **Robust realized integrated beta estimator with application to dynamic analysis of integrated beta**
Presenter: **Minseog Oh**, KAIST, Korea, South
Co-authors: Donggyu Kim, Yazhen Wang

A robust non-parametric realized integrated beta estimator is developed using high-frequency financial data contaminated by microstructure noises, which is robust to the stylized features, such as the time-varying beta and the dependence structure of microstructure noises. With this robust realized integrated beta estimator, dynamic structures of integrated betas are investigated, and an auto-regressive-moving-average (ARMA) structure is found. The ARMA model for daily integrated market betas is utilized to model this dynamic structure. This is called the dynamic realized beta (DR Beta). Further, a high-frequency data-generating process is introduced by filling the gap between the high-frequency-based non-parametric estimator and low-frequency dynamic structure. Then, a quasi-likelihood procedure for estimating the model parameters with the robust realized integrated beta estimator as the proxy is proposed. Asymptotic theorems are also established for the proposed estimator and conduct a simulation study to check the performance of finite samples of the estimator. The empirical study with the S&P 500 index and the top 50 large trading volume stocks from the S&P 500 illustrates that the proposed DR Beta model with the robust realized beta estimator effectively accounts for dynamics in the market beta of individual stocks and better predicts future market betas.

E1024: **Forecasting returns and optimizing global portfolios with machine learning: The Korean and U.S. stock markets**
Presenter: **Dohyun Chun**, Kangwon National University, Korea, South

The purpose is to evaluate the performance of international asset allocation strategies based on predictions of foreign exchange rates and stock market returns. Various machine learning models and a wide range of economic and financial variables are utilized to predict the KRW-USD exchange rate and U.S. and Korean stock market returns. The findings suggest that machine learning models outperform benchmark models in predicting both the exchange rate and stock market returns. Furthermore, a machine learning-driven global portfolio that accounts for exchange rate fluctuations demonstrates enhanced performance. Empirical evidence substantiating the use of machine learning techniques is presented to forecast foreign exchange rates and construct a compelling global portfolio.

EO038 Room 705 RECENT ADVANCES IN CLUSTERING AND HIGH DIMENSIONAL DATA ANALYSIS Chair: Sanjeena Dang

E0787: **Clustering high-dimensional count data**
Presenter: **Sanjeena Dang**, Carleton University, Canada
Co-authors: Andrea Payne, Anjali Silva, Steven Rothstein, Paul McNicholas

Multivariate count data are commonly encountered in bioinformatics. Although the Poisson distribution seems a natural fit for these count data, its multivariate extension is computationally expensive. Recently, mixtures of multivariate Poisson lognormal (MPLN) models have been used to efficiently analyze these multivariate count measurements. In the MPLN model, the counts, conditional on the latent variable, are modelled using a Poisson distribution, and the latent variable comes from a multivariate Gaussian distribution. Due to this hierarchical structure, the MPLN model can account for over-dispersion as opposed to the traditional Poisson distribution and allows for correlation between the variables. The mixture of multivariate Poisson-log normal distributions for high dimensional data is extended by incorporating a factor analyzer structure in the latent space. A family of parsimonious mixtures of multivariate Poisson lognormal distributions are proposed by decomposing the covariance matrix and imposing constraints on these decompositions. The performance of the model is demonstrated using simulated and real datasets.

E0795: **Weighted residual empirical processes, martingale transformations, and model checking for regressions**
Presenter: **Falong Tan**, Hunan University, China

A new methodology is proposed for testing the parametric forms of the mean and variance functions based on weighted residual empirical processes and their martingale transformations in regression models. The dimensions of the parameter vectors can be divergent as the sample size goes to infinity. The convergence of weighted residual empirical processes and their martingale transformation under the null and alternative hypotheses in the diverging dimension setting are then studied. The proposed tests based on weighted residual empirical processes can detect local alternatives distinct from the null at the fastest possible rate of order $n^1/2$ but are not asymptotically distribution-free. While the tests based on martingale-transformed weighted residual empirical processes can be asymptotically distribution-free, yet, unexpectedly can only detect the local alternatives converging to the null at a much slower rate of order $n^1/4$, which is somewhat different from existing asymptotically distribution-free tests based on martingale transformations. As the tests based on the residual empirical process are not distribution-free, a smooth residual bootstrap and verify the validity of its approximation in diverging dimension settings. Simulation studies and real data examples are conducted to illustrate the effectiveness of the tests.

E0794: **Recent developments in using mixtures of multivariate asymmetric distributions for classification**
Presenter: **Brian Franczak**, MacEwan University, Canada

Classification is defined as the process of assigning group labels to unlabelled observations. Classification is performed in multiple ways; for example, one can utilize unsupervised, semi-supervised, or fully supervised techniques when a finite mixture model is used for classification, called process model-based classification. Recent advances in the development of mixtures with multivariate density function capable of modelling skewness directly are discussed. In the context of classification applications, topics may include one or more strategies for handling data with missing values, working with high-dimensional data sets, rectifying issues with typical parameter estimation schemes, parameterizing tail-weight separately in each dimension of the data, or accounting for outlying or spurious observations. The proposed models will be demonstrated using both simulated and real data. Model performance will be assessed using standard metrics and by comparison to popular publicly available methods.

EO040 Room 708 STATISTICAL METHODS FOR FUNCTIONAL OBSERVATIONS Chair: Ci-Ren Jiang

E0262: **Interpretable discriminant analysis for functional data supported on random non-linear domains**
Presenter: **Eardi Lila**, University of Washington, United States

A novel framework for the classification of functional data supported on non-linear and possibly random manifold domains is introduced. The motivating application is the identification of subjects with Alzheimer's disease from their cortical surface geometry and associated cortical thickness map. The proposed model is based upon a reformulation of the classification problem as a regularized multivariate functional linear regression model. This allows us to adopt a direct approach to the estimation of the most discriminant direction while controlling for its complexity with appropriate differential regularization. The proposed method is applied to a pooled dataset from the Alzheimer's Disease Neuroimaging Initiative and the Parkinson's Progression Markers Initiative and is able to estimate discriminant directions that capture both cortical geometric and thickness predictive features of Alzheimer's Disease.

E0472: **Truncated estimation in functional generalized linear regression models**
Presenter: **Alexander Petersen**, Brigham Young University, United States
Co-authors: Xi Liu

Functional generalized linear models investigate the effect of functional predictors on a scalar response. An interesting case is when the functional predictor is thought to exert an influence on the conditional mean of the response only through its values up to a certain point in the domain. In the literature, models with this type of restriction on the functional effect have been termed truncated or historical regression models. A penalized likelihood estimator is formulated by combining a structured variable selection method with a localized B-spline expansion of the regression coefficient function. In addition to a smoothing penalty typical for functional regression, a nested group lasso penalty is also included, which guarantees the sequential entering of B-splines and thus induces the desired truncation on the estimator. An optimization scheme is developed to compute the solution path efficiently when varying the truncation tuning parameter. The convergence rate of the coefficient function estimator and consistency of the truncation point estimator is given under suitable smoothness assumptions. The proposed method is demonstrated through simulations and an application involving the effects of blood pressure values in patients who suffered a spontaneous intracerebral haemorrhage.

E0651: **Eigen-adjusted functional principal component analysis**
Presenter: **Ci-Ren Jiang**, National Taiwan University, Taiwan
Co-authors: Eardi Lila, Jane-Ling Wang, John Aston

Functional Principal Component Analysis (FPCA) has become a widely-used dimension reduction tool for functional data analysis. When additional covariates are available, existing FPCA models integrate them either in the mean function or in both the mean function and the covariance function. However, methods of the first kind are unsuitable for data that display second-order variation, while those of the second kind are time-consuming and make it difficult to perform subsequent statistical analyses on the dimension-reduced representations. To tackle these issues, an eigen-adjusted FPCA model is introduced that integrates covariates in the covariance function only through its eigenvalues. In particular, different structures on the covariate-specific eigenvalues corresponding to different practical problems are discussed to illustrate the model's flexibility and utility. To handle functional observations under different sampling schemes, local linear smoother to estimate the mean function and the pooled covariance function and a weighted least square approach to estimate the covariate-specific eigenvalues are employed. The convergence rates of the proposed estimators are further investigated under the different sampling schemes. In addition to simulation studies, the proposed model is applied to functional Magnetic Resonance Imaging scans collected within the Human Connectome Project for functional connectivity investigation.

EO060 Room 709 NEW CHALLENGES AND INSIGHTS IN HIGH-DIMENSIONAL STATISTICS Chair: Wei Luo

E0445: **Sparse convoluted rank regression in high dimensions**
Presenter: **Le Zhou**, Hong Kong Baptist University, Hong Kong
Co-authors: Boxiang Wang, Hui Zou

High-dimensional sparse penalized rank regression was studied in 2020, and it was shown to enjoy nice theoretical properties. Compared with the least squares, rank regression can substantially gain estimation efficiency while maintaining a minimal relative efficiency of 86.4%. However, the computation of penalized rank regression can be very challenging for high-dimensional data due to the highly nonsmooth rank regression loss. Convoluted rank regression is proposed, and the sparse penalized convoluted rank regression (CRR) for high-dimensional data is studied. Some interesting asymptotic properties of CRR are proven. Under the same key assumptions for sparse rank regression, the rate of convergence of the ł$_1$-penalized CRR for a tuning-free penalization parameter is established, and the strong oracle property of the folded concave penalized CRR is proven. Further, a high-dimensional Bayesian information criterion is proposed for selecting the penalization parameter in folded concave penalized CRR and its selection consistency is proven. An efficient algorithm is derived for solving sparse convoluted rank regression that scales well with high dimensions. Numerical examples demonstrate the promising performance of the sparse convoluted rank regression over the sparse rank regression. The theoretical and numerical results suggest that sparse convoluted rank regression enjoys the best of both sparse least squares regression and sparse rank regression.

E0552: **New tests for high-dimensional two-sample mean problems with consideration of correlation structure**
Presenter: **Songshan Yang**, Renmin University of China, China

A test statistic is proposed for two sample mean testing problems for high dimensional data by assuming the linear structure on high dimensional precision matrices. A new precision matrix estimation method considering its linear structure is first proposed, and the regularization method is implemented to select the true basis matrices that can further reduce the approximation error. Then the test statistic is constructed by imposing the estimation of the precision matrix. The proposed test is valid for both the low- and high-dimensional settings, even if the dimension of the data is greater than the sample size. The limiting null distributions of the proposed test statistic are derived under both null and alternative distributions. Extensive simulations are conducted to estimate the precision matrix and test the difference of the high-dimensional mean vector. Simulation results show that the proposed estimation method enjoys low estimation error for the precision matrix, and the regularization method is able to select the important basis matrix efficiently. The testing method performs well compared to existing methods, especially when the vector elements have unequal variances. A real data example is then provided to demonstrate the potential of the proposed method in real-world applications.

E0871: **Causal structural learning and application in epidemiology**
Presenter: **Changcheng Li**, Dalian University of Technology, China

The Population-based HIV Impact Assessment (PHIA) is an ongoing project that conducts nationally representative HIV-focused surveys for measuring national and regional progress toward UNAIDS 90-90-90 targets, the primary strategy to end the HIV epidemic. The PHIA survey offers a unique opportunity to better understand the key factors that drive the HIV epidemics in the most affected countries in sub-Saharan Africa. A novel causal structural learning algorithm is proposed to discover important covariates and potential causal pathways for 90-90-90 targets. Existing constrained-based causal structural learning algorithms are quite aggressive in edge removal. The proposed algorithm preserves more information about important features and potential causal pathways. It is applied to the Malawi PHIA (MPHIA) data set and leads to interesting results. The proposed algorithm is further compared and validated using BIC and using Monte Carlo simulations, and it is shown that it achieves improvement in true positive rates in important feature discovery over existing algorithms.

EC322 Room 04 CENSORED DATA Chair: Takeshi Emura

E0469: **Rank-based regression for doubly interval-censored data**
Presenter: **Seohyeon Park**, Korea University, Korea, South
Co-authors: Sangbum Choi, Zhezhen Jin, Wenbin Lu

In many biomedical fields, especially in studies of disease progressions, two sequential events where both event times tend are frequently encountered to be interval-censored due to regular examinations. Such a structure is called doubly interval censoring (DIC), and the primary interest is the elapsed time between two consecutive events. A weighted rank regression approach is proposed for DIC data under the semiparametric accelerated failure time model. After transforming DIC data into simple interval-censored data where true elapsed times may lie, estimating procedures are developed with a gehan-type weight by gathering all comparable pairs of observed residuals from transformed data. Moreover, it is generalized with data-dependent weights and is extended to clustered DIC data where the cluster size is potentially informative. An efficient resampling technique for the variance estimation is considered. Asymptotic properties are established, and numerical studies are conducted to demonstrate finite sample performances. Finally, the method with dental data is illustrated from the Signal Tandmobiel study to examine the effect of covariates on time to caries of four permanent first molars.

E1135: **Single-index mixture cure models: An application to a study of cardiotoxicity in breast cancer patients**
Presenter: **Ricardo Cao**, University of Coruna, Spain
Co-authors: Ana Lopez-Cheda, Beatriz Pineiro-Lamas

Standard survival models assume that the event of interest would always happen if there was sufficient follow-up time. However, this is not always realistic. For instance, HER2-positive breast cancer patients usually receive trastuzumab. Although this therapy has antitumor efficacy, it can cause a problem in the heart, known as cardiotoxicity, in some patients. In this context, there will be a fraction of individuals that will never suffer the side effect just because they are not susceptible to it. They are said to be cured in the sense that no matter how long you observe them, they will never experience the final event. To study the time until the cardiotoxicity appears, mixture cure models are appropriate. They allow us to estimate both the probability of being cured and the survival function of the uncured population, depending on some covariates. In the literature, nonparametric estimation of both functions is limited to continuous unidimensional covariates. This important gap is filled by considering multidimensional covariates and proposing a single-index model for dimension reduction. BC-Cardiotox, a dataset related to cardiotoxicity from the University Hospital of A Corua, is constructed and analyzed considering these techniques.

E1155: **Design on three-arm noninferiority trials using parametric and semiparametric methods for censored survival data**

Presenter: **Yi-Kang Tseng**, National Central University, Taiwan

A gold standard design for three-arm noninferiority trials, which includes an experimental group, a reference group and a placebo group using parametric and semiparametric models, was investigated for censored survival outcomes. In the literature so far, the sample size determination for a three-arm design with survival endpoints is limited. Therefore, to complement the literature, the purpose is to develop a statistical method to calculate the sample size for three-arm inferiority trials. Four parametric models (lognormal, Gompertz, Rayleigh, and log-logistic) and three semiparametric models (Cox, accelerated failure time and transformation models) were considered for survival data. The three groups' minimum required sample size and testing power with optimal allocation were considered the performance criteria for both parametric and semiparametric methods. Simulation studies were conducted to find out the performance of all methods. In particular, the impact of model misspecification was studied. The proposed procedure was finally applied to bladder cancer data.

Thursday 03.08.2023 15:25 - 17:05 Parallel Session O – EcoSta2023

EV274 Room Virtual R02 MODELLING AND FORECASTING Chair: Esra Kurum

E1005: **Forecasting economic activity with a neural network in uncertain times: Application to German GDP**
Presenter: **Boris Kozyrev**, Halle Institute for Economic Research (IWH), Germany
Co-authors: Oliver Holtemoeller

The forecasting and nowcasting performance of a generalized regression neural network (GRNN) is analyzed. First, evidence from Monte Carlo simulations for the relative forecast performance of GRNN depending on the true but unknown data-generating process is provided. The analysis shows that GRNN outperforms autoregressive-moving average models in various practical scenarios. An additional check of fitting ARMA using simulated samples is provided. As a result, even though they yield similar to GRNN predictions in many cases, existing ARMA fitting approaches often cannot properly identify a true DGP. Later, GRNN is applied to forecast quarterly German GDP growth, distinguishing between "normal" times and situations with significantly different time-series behaviour, such as during the COVID recession and recovery. The specific data transformation needs to be implemented, i.e., dividing aggregated level values of each indicator by the corresponding GDP value. Then, these ratios are used to perform one-step-ahead forecasting using GRNN. After that, a set of GDP nowcasts is obtained using actual aggregated observations within a given quarter. This algorithm has a high forecasting power, outperforming traditional nowcasting models (AR(1), DFM, model averaging), especially during the COVID-19 crisis.

E1094: **Step by step: A quarterly evaluation of EU commissions' GDP forecasts**
Presenter: **Katja Heinisch**, Halle Institute for Economic Research, Germany

Annual growth forecasts by the European Commission are important figures for policy-making and provide a benchmark for many forecasters. However, they are usually based on quarterly estimates, which are hardly known and do not get much attention. Therefore, a detailed analysis is provided for the multi-period ahead quarterly GDP growth forecasts for the EU, euro area, and several EU member states with respect to first-release and current-release data. Forecast revisions and forecast errors are analyzed, and the results show that the forecasts are not systematically biased. However, a significant overestimation of short-time horizons is identified for several member states. The highest performance is not achieved for the current quarter for all countries, although a high forecast revision occurs in the last step (from one-quarter-ahead forecast to the current quarter). Furthermore, the final forecast revision in the current quarter is generally downward biased for almost all countries. Overall, the differences in mean forecast errors are minor when using real-time data or pseudo-real-time data. The forecast performance also varies across countries, with smaller countries and Central and Eastern European countries (CEEC) having larger forecast errors. Evidence is provided that there is still room for improvement in forecasting techniques both for nowcasts but also forecasts up to 8 quarters ahead.

E1136: **A term structure dynamic model with correlated residuals: A comparative analysis**
Presenter: **Antonella Congedi**, University of Salento, Italy
Co-authors: Sandra De Iaco, Sabrina Maggio

Recently, modelling the term structure of interest rates has gained particular relevance in the economic and financial literature and has become one of the main research topics. Nowadays, the growth of financial markets and emerging derivative instruments require the development of new techniques for estimating and forecasting interest rates that could be adapted to reality. Term structure modelling should consider two significant dimensions, time and maturity, although, in the literature, these were treated separately or analyzed through multivariate techniques. An alternative geostatistical approach was based on the use of a correlation function which was assumed to be dependent only on the temporal lag. Differently from the existing contributions, the aim is to propose a geostatistical model for the term structure of spot rates, where the joint evolution concerning time and maturity is considered. In particular, three hypotheses on the trend component of the random field are assumed: i) constant, ii) dependent only on time, and iii) dependent on time and maturity. Finally, a comparison among the predictive performance of models in that different hypotheses is proposed.

E0296: **Some first results from an agent-based model of consumer demand**
Presenter: **Georgios Alkis Tsiatsios**, National and Kapodistrian University of Athens, Greece
Co-authors: Iraklis Kollias, Evangelos Melas, John Leventides, Costas Poulios

An agent-based model of consumer choice is studied. One of the primary advantages of this type of modelling is that it allows for various forms of heterogeneity between agents and different behavioural rules regarding decision-making. Here, agents are heterogeneous concerning their preferences and income. The behavioural rule is utility maximization. Commodity prices are subject to different states of the world. Hence, the model is conducive to the presence of risky outcomes. Because agent-based models are based in large part on simulations, the modeller can derive relations of a dynamic sort that are otherwise unobtainable within the framework of the representative agent model of consumer choice and classical demand theory. An example of this is that Walrasian demand now is not simply a point in commodity space but rather a complete trajectory for each agent or each group of agents. The agent-based data-driven model is used to derive demand functions of a heterogeneous set of consumers of a market consisting of different types of products. The demand functions derived are multiparametric. Graphically, various forms are depicted, conclusions are drawn about the market, and the economy is simulated.

EO017 Room 02 TEXT DATA Chair: Ana Colubi

E1312: **HiTEc: Exploiting text data in applications**
Presenter: **Ana Colubi**, Kings College London, Cyprus
Co-authors: Louisa Kontoghiorghes

HiTEc is a COST Action integrating cutting-edge analytic developments involving innovative sources of information, such as text, functions, perceptions or imprecise data, in econometrics. Several applications within this topic will be discussed. As a meta-result and proof-of-concept, a text-based indicator of the success of a COST Action will be described. The indicator is based on topic modelling. Specifically, it uses Latent Dirichlet Allocation (LDA) as the primary tool. Additional properties and uses of the indicator will also be outlined.

E1174: **Measuring and comparing the thematic prevalence using a parametric and distribution-free bootstrap two-sample test**
Presenter: **Louisa Kontoghiorghes**, Kings College London, United Kingdom
Co-authors: Ana Colubi

An approach has been introduced to measure the prevalence of specific subjects in a corpus using keywords. Instead of estimating the frequencies of the keywords directly, the method utilizes topic modelling to extract the structure of the topics within the documents. This enables the computation of the subject prevalence by averaging the frequencies of the keywords within the topics while taking into account the importance of the topics in the documents. Using a distribution-free bootstrap, a hypothesis test comparing the keyword-based prevalence has been proposed. An alternative parametric bootstrap test is proposed and compared to the existing test. It is applied to the sentiment analysis of Jane Austen's novels to demonstrate the methodology.

E1212: **Predicting ICU readmission with a hybrid BERTopic-LSTM approach on electronic health records**
Presenter: **Ling-Jing Kao**, National Taipei University of Technology, Taiwan

Co-authors: Chih-Chou Chiu, Chung-Min Wu, Te-Nien Chien, Chengcheng Li

The high incidence of ICU readmissions poses a significant challenge in healthcare, resulting in increased expenses and suboptimal patient outcomes. A novel method is presented for predicting ICU readmission from electronic health records (EHRs) using a hybrid BERTopic and Long Short-Term Memory (LSTM) network approach. The model integrates the benefits of unsupervised topic modelling with supervised deep learning to effectively capture the intricate associations between patient attributes and readmission risk. A dataset of 5,000 ICU patient records was leveraged, where BERTopic was initially employed to cluster patients based on their EHRs. Then a supervised LSTM network was utilized for training on the clustered data, incorporating both the EHRs and patient demographics as inputs to forecast readmission risk. The method outperforms existing readmission prediction models based on traditional machine learning approaches, yielding an AUC-ROC of 0.80. Additionally, it is demonstrated that the proposed model can identify key risk factors for readmission, including comorbidities, length of stay, and ICU admission type. In summary, the aim is to highlight the efficacy of a hybrid BERTopic and LSTM network approach for predicting ICU readmission from EHRs. This approach holds promise in enhancing patient outcomes and reducing healthcare costs by enabling early intervention and targeted care management.

EO132 Room Virtual R01 EXTREME RISK MEASURES ESTIMATION IN VARIOUS ATTRACTION DOMAINS Chair: Antoine Usseglio-Carleve

E0621: **Extreme expectile estimation for short-tailed data**

Presenter: **Abdelaati Daouia**, Fondation Jean-Jacques Laffont, France

Co-authors: Simone Padoan, Gilles Stupfler

The use of expectiles in risk management has recently gathered remarkable momentum due to their excellent axiomatic and probabilistic properties. In particular, the class of elicitable law-invariant coherent risk measures only consists of expectiles. While the theory of expectile estimation at central levels is substantial, tail estimation at extreme levels has so far only been considered when the tail of the underlying distribution is heavy. This is the first work to handle the short-tailed setting where the loss (*, e.g.* negative log-returns) distribution of interest is bounded to the right, and the corresponding extreme value index is negative. An asymptotic expansion of tail expectiles is derived in this challenging context under a general second-order extreme value condition, which allows to come up with two semiparametric estimators of extreme expectiles with their asymptotic properties in a general model of strictly stationary but weakly dependent observations. A simulation study and a real data analysis from a forecasting perspective are performed to verify and compare the proposed competing estimation procedures.

E0653: **Analysis of variability in extremes**

Presenter: **Chen Yan**, INRAE/Inria, France

Co-authors: Stephane Girard, Thomas Opitz, Antoine Usseglio-Carleve

ANOVA is a widely used statistical technique to compare the means of several groups of independently sampled data. However, examining tail behaviour instead of the mean is more relevant in some cases. Therefore, ANOVEX (ANalysis Of Variability in EXtremes), an analogue of ANOVA, is proposed to compare the behaviour of extremes across J>1 groups. The ANOVEX test involves selecting a number L>1 of extreme quantiles within each group, estimated using methods such as the Weissman estimator. Within-group and between-group variances of extreme log quantiles are calculated. It is demonstrated that under the null hypothesis of the same behaviour across groups, the ratio of these variances converges to a chi-square distribution with J-1 degrees of freedom after normalization. To further enhance the applicability of ANOVEX, it is proposed to combine it with a decision tree algorithm for clustering of extremes, where each observation comes with K covariates. At each tree node, the ANOVEX test is applied for all possible splits of all covariates on data belonging to that node. The most significant test from ANOVEX determines the best splitting rule of the node. Once a large tree is built, pruning by fusing two leaves is applied if the best test statistic of their common parent is not significant. This ANOVEX-tree algorithm is applied to examples with K=1 covariate and a real data example of wildfire-burnt areas in the US with more than 500,000 samples and over 30 covariates.

E0711: **A refined extreme quantiles estimator for Weibull tail-distributions**

Presenter: **Jonathan El Methni**, Universite Paris Cite, France

Co-authors: Stephane Girard

The problem of extreme quantiles estimation is addressed for Weibull tail distributions. Since such quantiles are asymptotically larger than the sample maxima, their estimation requires extrapolation methods. In the case of Weibull tail distributions, classical extreme-value estimators are numerically outperformed by estimators dedicated to this set of light-tailed distributions. The latter estimators of extreme quantiles are based on two estimators: an order statistic to estimate an intermediate quantile and an estimator of the Weibull tail coefficient. The common practice is to select the same intermediate sequence for both estimators. The aim is to show how an adapted choice of two different intermediate sequences leads to a reduction of the asymptotic bias associated with the resulting refined estimator. The asymptotic normality of the refined estimator is established, and a data-driven method is introduced for the practical selection of the intermediate sequences. The approach is compared to three estimators of extreme quantiles dedicated to Weibull tail distributions in a simulation study. An illustration of real data is also provided.

E0834: **Bias- and variance-corrected asymptotic Gaussian inference about extreme expectiles**

Presenter: **Antoine Usseglio-Carleve**, Avignon Universita, France

Co-authors: Abdelaati Daouia, Gilles Stupfler

The executile is a prime candidate for being a standard risk measure in actuarial and financial contexts for its ability to recover information about probabilities and typical behaviour of extreme values as well as its excellent axiomatic properties. A series of recent papers have focused on executile estimation at extreme levels, with a view to gathering essential information about low-probability. These high-impact events are of most interest to risk managers. Actual inference about extreme executiles is a difficult question, however, due to their least squares formulation making them very sensitive to tail heaviness, even though the obtention of accurate confidence intervals is paramount if the expectile risk measure is to be used in practical applications. This article focuses on asymptotic Gaussian inference about tail expectiles in the challenging context of heavy-tailed observations. An in-depth analysis of the proofs of asymptotic normality results is used for two classes of extreme expectile estimators to derive bias- and variance-corrected Gaussian confidence intervals. Unlike previous attempts in the literature, these are well-rooted in statistical theory and can accommodate underlying distributions that display a wide range of tail behaviours. A large-scale simulation study and real data analyses confirm the versatility of the proposed technique.

EO218 Room 603 FORECAST COMBINATION Chair: Andrey Vasnev

E0452: **Global combinations of expert forecasts**

Presenter: **Ryan Thompson**, University of New South Wales, Australia

Co-authors: Andrey Vasnev, Yilin Qian

Expert forecast combination- aggregating individual forecasts from multiple subject-matter experts- is a proven approach to economic forecasting. To date, research in this area has exclusively concentrated on local combination methods, which handle separate but related forecasting tasks in isolation. Yet, the machine learning community has known for over two decades that global methods, which exploit task-relatedness, can improve on local methods that ignore it. Motivated by the possibility for improvement, a framework is introduced for globally combining expert forecasts. Through this framework, global versions of several existing forecast combinations are developed. To evaluate the efficacy of these new global forecast combinations, extensive comparisons are reported using synthetic and real data. Our real data comparisons, which involve expert forecasts of core economic indicators in the Eurozone, are the first empirical evidence that the accuracy of global combinations of expert forecasts can surpass local combinations.

E0820: **Factor models and forecast combinations**
Presenter: **Rachida Ouysse**, University of New South Wales, Australia
Co-authors: Andrey Vasnev

Principal components (PC) forecasts are more accurate than many single econometric models. However, the principal components are computed from the predictors without accounting for their relationship with the forecast target variables. An alternative approach is the factor combination of forecasts. An alternative scenario is considered where multiple forecasters have access to subsets (possibly overlapping) of the full information set. Forecasters use a single-factor model to construct factor forecasts. The performance of combining these partial information forecasts with a single factor forecast from the full set of predictors is analysed. Combination methods include equal weight, optimal weight, shrinkage weights, and principal component combination are considered. An application to forecasting the monthly growth rate of U.S. industrial production, where the full set of predictors consists of 130 economic indicators, shows that combining forecasts outperforms the full information factor forecasts. The Shrinkage weights combination performed better than equal weights. More data is not better for forecast performance than a weighted combination of selective consensus from its snippets.

E0655: **Optimal model averaging for single-index models with divergent dimensions**
Presenter: **Wendun Wang**, Erasmus University Rotterdam, Netherlands

A new approach is offered to address the model uncertainty in (potentially) divergent-dimensional single-index models (SIMs). A model-averaging estimator based on cross-validation, which allows the dimension of covariates, and the number of candidate models to increase with the sample size, is proposed. It is shown that when all candidate models are misspecified, the model-averaging estimator is asymptotically optimal, with its squared loss asymptotically identical to that of the infeasible best possible averaging estimator. In a different situation where correct models are available in the model set, the proposed method assigns all weights to the correct models asymptotically. Averaging regularized estimators and prescreening methods to deal with high-dimensional covariates are also proposed. The method via simulations and an empirical application are illustrated.

E0477: **Forecast combination puzzle in the HAR model**
Presenter: **Andrey Vasnev**, University of Sydney, Australia
Co-authors: Adam Clements

Given its simplicity and consistent empirical performance, the Heterogeneous Autoregressive (HAR) model of Corsi has become the benchmark model for predicting realized volatility. Many modifications and extensions to the original model have been proposed that often only provide incremental forecast improvements. A step back is taken, and the HAR model is viewed as a forecast combination that combines three predictors: previous day realization (or random walk forecast), previous week average, and previous month average. When apply- ing the Ordinary Least Squares (OLS) to combine the predictors, the HAR model uses optimal weights that are known to be problematic in the forecast combination literature. An average of simpler HAR-style models and a simple average forecast often outperforms the optimal combination in many empirical applications. The performance of these simple combination forecasts is investigated for the realized volatility of the Dow Jones Industrial Average equity index and a sample of individual constituent stocks, as well as across a range of other as- sets, commodities, exchange rates and a range of global equity market indices. In all cases, dramatic improvements in forecast accuracy across all horizons and different time periods are found. This is the first time the forecast combination puzzle has been identified in this context.

EO080 Room 604 RECENT DEVELOPMENTS IN TIME SERIES ECONOMETRICS Chair: Cy Sin

E0726: **Risk-return trade-off in the Bitcoin market: Downside risk and sentiment**
Presenter: **Yin-Feng Gau**, National Central University, Taiwan
Co-authors: Hann Chang

Extending the Heterogeneous Autoregressive Model of Realized Volatility (HAR-RV), the upside and downside risks of the Bitcoin returns in gauging the risk-return relationship in the Bitcoin market are estimated. Using the intraday data of Bitcoin prices from January 2018 to September 2022, the aim is to study how the risk-return relationship of Bitcoin relates to investor sentiment and risk shifts during the COVID-19 pandemic. The results show a significant and positive relationship between the downside risk and returns in the Bitcoin market, implying investors require higher returns to compensate for the downside risks. The positive relationship between returns and downside risk is even stronger during the COVID-19 pandemic.

E0583: **Binary choice models with multiple integrated predictors**
Presenter: **Hsein Kew**, Monash University, Australia

A binary probit model with multiple integrated predictors is considered. This model is useful for predicting a binary recession using interest rates on long and short-maturity debt and private debt interest rates with different degrees of default risk. A constrained non-linear least squares estimator is considered to estimate the model's unknown parameters. Monte Carlo study shows that this estimator produces estimates with better precision for a relatively small sample size than an unconstrained, non-linear least squares estimator. This model is applied to forecast U.S. recessions, and the forecast performance is compared with binary probit models with stationary predictors in terms of in-sample and out-of-sample predictive power.

E0736: **Prediction for multivariate time series models with deterministic time trends**
Presenter: **YanShuo Pan**, National Tsinghua University, Taiwan
Co-authors: Ching-Kang Ing, Cy Sin

When model selection in time series data is referred to, most existing literature considers time series models with a constant mean, while time series data containing deterministic time trends (DTTs) are becoming more common. A multivariate version of the Misspecification-Resistant Information Criterion (MRIC) is proposed for model selection in time series data with DTTS. Unlike the conventional model selection methods, such as Bayesian Information Criterion (BIC) and Akaike Information Criterion (AIC), which focus on correctly specified models, the MRIC proposed by other researchers is designed to cover misspecified models. An asymptotic expansion of the mean squared prediction error (MSPE) is derived in misspecified time series models with DTTs, building on previous work, which provided an asymptotic expression for the MSPE of the least squares predictor. The aim is to show that the multivariate MRIC (MMRIC) achieves asymptotic efficiency regardless of whether the true model is among the candidate models or not. Furthermore, MMRIC can be applied to choose the best multi-step predictive model, which is important for practical applications of time series data.

E0809: **Re-balancing hedge position with statistics of hedge ratios: Concepts and applications**
Presenter: **Cy Sin**, National Tsing Hua University, Taiwan

Recently, strong evidence has been claimed to be found indicating that advanced econometric models do not improve hedge efficiency significantly, if at all. As a matter of fact, dynamic hedging attempts to strike a balance between hedging effectiveness and transaction costs. Using the Garch asymptotic theories, the asymptotic properties of the hedge ratio are derived. As a result, a natural and simple statistic of re-balancing is constructed, namely, the (asymptotic) standard deviation of the hedge ratio. The method is applied to a number of paired variables, such as WTI Crude Oil Futures and Spot Price.

EO147 Room 605 PROBABILITY AND STOCHASTIC GEOMETRY WITH STATISTICAL APPLICATIONS Chair: Claudio Durastanti

E0556: **Detecting a late changepoint in a growing network**
Presenter: **Gianmarco Bet**, Universita degli Studi di Firenze, Italy

Motivated by the problem of detecting a change in the evolution of a network, the preferential attachment random graph model with a time-dependent attachment function is considered. Our goal is to detect whether the attachment mechanism changes over time based on a single network snapshot. This question is cast as a hypothesis testing problem, where the null hypothesis is a preferential attachment model with a constant affine attachment parameter δ_0, and the alternative hypothesis is a preferential attachment model where the affine attachment parameter changes from δ_0 to δ_1 at an unknown changepoint time τ_n. It is focused on the regime where δ_0 and δ_1 are fixed, and the changepoint occurs close to the observation time n of the network (i.e., $\tau_n = n - cn^\gamma$ with $c > 0$ and $\gamma \in (0,1)$). This corresponds to the relevant scenario where the aim is to detect the changepoint shortly after it has happened. Two tests based on the number of vertices with a minimal degree are presented, showing that these are asymptotically powerful when $\gamma > 1/2$. The first test requires knowledge of δ_0. The second test is significantly more involved and does not require the knowledge of δ_0 while achieving the same performance guarantees. It is proved that the test statistics for both tests are asymptotically normal, allowing for accurate calibration of the tests.

E0592: **The volume of random Beta polytopes in high dimensions**
Presenter: **Nicola Turchi**, University of Milano-Bicocca, Italy
Co-authors: Gilles Bonnet, Zakhar Kabluchko

Beta polytopes are a class of random polytopes which arise as convex hulls of independent random points distributed according to a certain radially-symmetric probability distribution supported on the Euclidean ball, called the beta distribution. As the space dimension grows, the expected fraction of the volume that these polytopes fill within their supporting balls can be asymptotically negligible or not, depending on the number of points picked in each dimension. An overview of how to quantify this statement is given, first showing a rough threshold for the aforementioned growth and, secondly, a more precise one, namely, how many points are needed to get any fraction in average.

E0628: **Lipschitz-Killing curvatures for arithmetic random waves**
Presenter: **Valentina Cammarota**, Sapienza University of Rome, Italy
Co-authors: Domenico Marinucci, Maurizia Rossi

The purpose is to show that the Lipschitz-Killing Curvatures for the excursion sets of Arithmetic Random Waves (toral Gaussian eigenfunctions) is dominated, in the high-frequency regime, by a single chaotic component. The latter can be written as a simple explicit function of the threshold parameter times the centred norm of these random fields; as a consequence, these geometric functionals are fully correlated in the high-energy limit. The derived formulae show a clear analogy with related results on the round unit sphere and suggest the existence of a general formula for geometric functionals of random eigenfunctions on Riemannian manifolds.

E0827: **Spherical Poisson waves**
Presenter: **Claudio Durastanti**, Sapienza University of Rome, Italy
Co-authors: Domenico Marinucci, Anna Paola Todino, Solesne Bourguin

A model of Poisson random waves is discussed, defined in the sphere, to study Quantitative Central Limit Theorems when both the rate of the Poisson process (that is, the expected number of the observations sampled at a fixed time) and the energy (i.e., frequency) of the waves (eigenfunctions) diverge to infinity. Finite-dimensional distributions, harmonic coefficients and convergence in law in functional spaces are considered, and the interplay is investigated carefully between the rates of divergence of eigenvalues and Poisson governing measures.

EO095 Room 606 ADVANCES IN COMPLEX TIME SERIES ANALYSIS Chair: Weilin Chen

E0199: **Sparse change detection in high-dimensional linear regression**
Presenter: **Tengyao Wang**, London School of Economics, United Kingdom

A new methodology, 'charcoal' for estimating the location of sparse changes in high-dimensional linear regression coefficients, without assuming that those coefficients are individually sparse, is introduced. The procedure works by constructing different sketches (projections) of the design matrix at each time point, where consecutive projection matrices differ in sign in exactly one column. The sequence of sketched design matrices is then compared against a single sketched response vector to form a sequence of test statistics whose behaviour shows a surprising link to the well-known CUSUM statistics of univariate changepoint analysis. Strong theoretical guarantees are derived for the estimation accuracy of the procedure, which is computationally attractive, and simulations confirm that the methods perform well in a broad class of settings.

E0358: **Rank and factor loadings estimation in time series tensor factor model by pre-averaging**
Presenter: **Weilin Chen**, London School of Economics and Political Science, United Kingdom
Co-authors: Clifford Lam

As a major dimension reduction tool, the idiosyncratic components of a tensor time series factor model can exhibit serial correlations, especially in financial and economic applications. This rules out a lot of state-of-the-art methods that assume white idiosyncratic components or even independent/Gaussian data. While the traditional higher-order orthogonal iteration (HOOI) is proved to be convergent to a set of factor loading matrices, the closeness of them to the true underlying factor loading matrices is, in general, not established, or only under i.i.d. Gaussian noises. Under the presence of serial and cross-correlations in the idiosyncratic components and time series variables with only bounded fourth-order moments, a pre-averaging method that accumulates information is proposed from tensor fibres for better estimating all the factor loading spaces. The estimated directions corresponding to the strongest factors are then used for projecting the data for a potentially improved re-estimation of the factor loading spaces themselves, with theoretical guarantees and rate of convergence spelt out. A new rank estimation method is also proposed which utilises correlation information from the projected data. Extensive simulations are performed and compared to other state-of-the-art or traditional alternatives. A set of matrix-valued portfolio return data is also analysed.

E0447: **Imputation for tensor time series**
Presenter: **Zetai Cen**, London School of Economics and Political Science, United Kingdom
Co-authors: Clifford Lam

It is prevalent to have missing data in different areas, such as econometrics, and imputation is one of the common approaches to deal with it. With the fast-growing data size and complicated data structure, tensor time series is considered, and an imputation procedure based on factor analysis is proposed. Other researchers generalised The idea previously, but more general data are allowed. Specifically, a re-arrangement algorithm is used to construct different blocks of tensor data, and previous work has been adopted to estimate the tensor factor model for imputation. Thus, weak factors and serial and cross-correlations in the idiosyncratic errors and time series variables are allowed with bounded fourth-order moments. Also, iterative imputation is performed by re-estimating the factor structure to improve the imputation result. Simulations under different settings are performed, with appealing results even with weak factors and heavy-tailed data, and the one-step iteration is good enough and enjoys a short imputation time.

E0484: **Adaptive wavelet domain principal component analysis for nonstationary time series**
Presenter: **Marina Knight**, University of York, United Kingdom

Co-authors: Matthew Nunes, Jessica Hargreaves

High-dimensional multivariate nonstationary time series, i.e. data whose second-order properties vary over time, are common in many scientific and industrial applications. A novel wavelet domain dimension reduction technique for nonstationary time series is proposed. By constructing a time-scale adaptive principal component analysis of the data, the proposed method is able to capture the salient dynamic features of the multivariate time series. A new time and scale-dependent cross-coherence measure are also introduced, and it is shown to successfully quantify the extent of association between a multivariate nonstationary time series and its proposed wavelet domain principal component representation. Theoretical results establish that the associated estimation scheme enjoys good bias and consistency properties when determining wavelet domain principal components of input data. The proposed method is illustrated using extensive simulations, and its applicability on a real-world dataset arising in a neuroscience study is demonstrated.

EO102 Room 702 REPEATED MEASURES, FDA, NONPARAMETRIC REGRESSION, AND REGULARIZED T-TEST Chair: Yuedong Wang

E0236: **Linear models for multivariate repeated measures data**
Presenter: **Anuradha Roy**, The University of Texas at San Antonio, United States
Co-authors: Timothy Opheim

Multivariate repeated measures data, where observations are made on p response variables, and each response variable is measured over n sites or time points, construct matrix-variate response variables and arise across a wide range of disciplines, including medical, environmental and agricultural studies. The popularity of the classical general linear model (CGLM) is primarily due to the ease of modelling and authentication of the appropriateness of the model. However, CGLM is not appropriate and thus not applicable for multivariate data with multiple doubly correlated measurements. Extending the linear model for these doubly correlated multivariate data is proposed. Maximum likelihood estimates of the intercept and slope matrix parameters are derived. The practical implications of the methodological aspects of the proposed extended model for multivariate repeated measures data are demonstrated using two medical datasets.

E0300: **Regularized t distribution: Definition, properties and applications**
Presenter: **Tiejun Tong**, Hong Kong Baptist University, Hong Kong

Omics data analysis plays an important role in biological research. An important task for gene expression data analysis is to identify genes that are differentially expressed between two or more groups. Nevertheless, as biological experiments are often measured with a relatively small number of samples, how accurately estimating the variances of gene expression becomes a challenging issue. To tackle this problem, a regularized t distribution is introduced, and its statistical properties are derived, including the probability density function and the moment generating function. The noncentral regularized t distribution is also introduced for computing the statistical power of hypothesis testing. For practical applications, the regularized t distribution is applied to establish the null distribution of the regularized t statistic and then formulate it as a regularized t-test for detecting the differentially expressed genes. Simulation studies and real data analysis show that the regularized t-test performs much better than the Bayesian t-test in the limma package, in particular when the sample sizes are small.

E1177: **Communication-efficient distributed portfolio selection strategy**
Presenter: **Hongmei Lin**, Shanghai University of International Business and Economics, China

Modern portfolio theory is one of the most influential basic theories in the field of financial investment. The development of successful portfolio selection strategies requires that the strategies have good out-of-sample performance with substantial rewards, provide diversification benefits with controllable risk, and are easy to operate and maintain. However, as the global financial market expands, a large number of different asset data are often collected from different sources with different times and locations, and the existing portfolio strategy cannot deal with this situation. A new estimation approach is provided for the portfolio selection strategy in a distributed system, and the theoretical results are established. Monte Carlo simulations are further applied to evaluate its finite sample performance, and the usefulness of the algorithm is also illustrated through the NYSE datasets.

E0297: **Optimal-*k* difference sequence in nonparametric regression**
Presenter: **Wenlin Dai**, Renmin University of China, China
Co-authors: Xingwei Tong, Tiejun Tong

Difference-based methods have been attracting increasing attention in nonparametric regression, particularly for estimating the residual variance. To implement the estimation, one needs to choose an appropriate difference sequence, mainly between the optimal difference sequence and the ordinary difference sequence. The difference sequence selection is a fundamental problem in nonparametric regression, and it has remained a controversial issue for over three decades. The aim is to tackle this challenging issue from a unique perspective, namely by introducing a new difference sequence called the optimal-*k* difference sequence. The new difference sequence not only provides a better balance between the bias-variance trade-off but also dramatically enlarges the existing family of difference sequences that includes the optimal and ordinary difference sequences as two important special cases. Further, it is demonstrated, by both theoretical and numerical studies, that the optimal-*k* difference sequence has been pushing the boundaries of the knowledge in difference-based methods in nonparametric regression, and it always performs the best in practical situations.

EO024 Room 703 STATISTICAL NETWORK ANALYSIS III Chair: Joshua Cape

E0485: **A strategic model of software dependency networks**
Presenter: **Angelo Mele**, Johns Hopkins University, United States
Co-authors: Co-Pierre Georg

Modern software development involves collaborative efforts and re-using existing software packages and libraries to reduce the cost of developing new software. However, package dependencies expose developers to the risk of contagion from bugs or other vulnerabilities. The aim is to study the formation of dependency networks among software packages and libraries, guided by a structural model of network formation with observable and unobservable heterogeneity. A package maintainer's costs, benefits and link externalities are estimated using a scalable algorithm and data from 1,131,342 dependencies of 17,081 packages of the Rust programming language. It is found evidence of a positive externality created by coders on other coders through the creation of dependencies. It is also found that homophily and competition motives coexist in creating the network.

E0780: **Lost in the shuffle: Testing power in the presence of errorful network vertex labels**
Presenter: **Vince Lyzinski**, University of Maryland, College Park, United States
Co-authors: Ayushi Saxena

Many two-sample network hypothesis testing methodologies operate under the implicit assumption that the vertex correspondence across networks is a priori known. Power degradation in two-sample graph hypothesis testing is considered when there are misaligned/label-shuffled vertices across networks. In the context of random dot product and stochastic block model networks, the power loss due is theoretically explored to shuffling for a pair of hypothesis tests based on Frobenius norm differences between estimated edge probability matrices or between adjacency matrices. The loss in testing power is further reinforced by numerous simulations and experiments, both in the stochastic block model and in the random dot product graph model, where the power loss is compared across multiple recently proposed tests in the literature. Lastly, the impact that shuffling can have in real-data testing is demonstrated in a pair of examples from neuroscience and social network analysis.

E0889: **Two generalizable strategies for scalable inference from network data**
Presenter: **Srijan Sengupta**, North Carolina State University, United States

Massive network data are becoming increasingly common in scientific applications. Existing community detection methods are computationally infeasible for such massive networks. Two generalizable strategies are proposed for scalable inference from network data: SONNET and predictive sketching. SONNET is a ssubsampling-baseddivide-and-conquer algorithm, where the original network is split into multiple subnetworks with a common overlap. Statistical inference is carried out for each subnetwork, and the results from individual subnetworks are aggregated by leveraging the overlap. The core idea of predictive sketching is to avoid large-scale matrix computations by breaking up the task into a more negligible, more minor computation plus a large number of vector computations which can be carried out in parallel. Under the proposed method, the inferential task of interest is carried out on a small subgraph to estimate the relevant model parameters. The remaining nodes are added one by one using only vector computations. These two strategies are applied to various inference tasks, such as community detection, parameter estimation, model selection, and hypothesis testing.

E0905: **Discovering underlying dynamics in time series of networks**
Presenter: **Avanti Athreya**, Johns Hopkins University, United States
Co-authors: Zachary Lubberts, Carey Priebe, Youngser Park

Understanding dramatic changes in the evolution of networks is central to statistical network inference. A joint network model has been considered in which each node has an associated time-varying low-dimensional latent vector of feature data, and connection probabilities are functions of these vectors. Under mild assumptions, the time-varying evolution of the constellation of latent vectors exhibits a low-dimensional manifold structure under a suitable notion of distance. This distance can be approximated by a measure of separation between the observed networks. Euclidean representations exist for the underlying network structure, characterized by this distance, at any given time. These Euclidean representations and their data-driven estimates permit the visualization of network evolution and transform network inference questions such as change-point and anomaly detection into a classical setting. The methodology is illustrated with real and synthetic data, and change points are identified corresponding to shifts in pandemic policies in a communication network of a large organization.

EO168 Room 705 RECENT ADVANCES IN HIGH-DIMENSIONAL AND DEPENDENT DATA ANALYSIS Chair: Ching-Kang Ing

E0257: **A negative moment bound for integrated autoregressions with polynomial time trend and its applications**
Presenter: **Shu-Hui Yu**, Institute of Statistics, Taiwan

A moment bound is established for the inverse of the normalized Fisher information matrix in an integrated autoregressive model with polynomial time trends. This bound serves as the primary technical tool to derive an asymptotic expression for the mean squared prediction error (MSPE) of the least squares predictor. The derived expression is noteworthy because it provides the first precise assessment of nonstationarity, model complexity, and model over-specification impacts on the MSPE, forming a solid theoretical foundation for certain model selection criteria.

E0540: **Feature selection for high-dimensional heteroscedastic regression models**
Presenter: **PoHsiang Peng**, National Tsing Hua University, Taiwan
Co-authors: HaiTang Chiou, Hsueh-Han Huang, Ching-Kang Ing

Feature selection for high-dimensional linear heteroscedastic models is considered. Inspired by the connection between the linear heteroscedastic function and the interaction model, a two-stage algorithm is designed to choose the relevant features in the aforementioned high-dimensional model. Moreover, when it is unknown whether the functional form of heteroscedasticity is linear or multiplicative, a data-driven method is provided to select between the two alternatives. The selection consistency of the proposed method is proved and its performance is illustrated via numerical simulations. Further, the method is applied to identify defective tools in the semiconductor manufacturing process.

E0833: **Importance weighted orthogonal greedy algorithm with estimated weight function**
Presenter: **Shinpei Imori**, Hiroshima University, Japan
Co-authors: Ching-Kang Ing

Greedy-type algorithms are feasible for prediction in high-dimensional linear regression models when the number of explanatory variables is larger than the sample size. A greedy algorithm is studied under the covariate shift, where the distribution of explanatory variables in training data possibly differs from that in test data. The proposed algorithm needs to use an unknown weight function based on the density ratio. A sufficient condition is given in order that the proposed algorithm with the estimated weight function archives a good convergence rate.

E0947: **Model selection for unit-root time series with many predictors**
Presenter: **Ching-Kang Ing**, National Tsing Hua University, Taiwan

Model selection is studied for a general unit-root time series when many predictors are present. A new model selection algorithm called FHTD is proposed that leverages the advantages of forward stepwise regression (FSR), the high-dimensional information criterion (HDIC), a backward elimination method based on HDIC, and a data-driven thresholding (DDT) approach. By deriving a new functional central limit theorem for multivariate linear processes, along with a uniform lower bound for the minimum eigenvalue of the sample covariance matrices of the series under study, the sure screening property of FSR and the selection consistency of FHTD under some mild assumptions that allow for unknown locations and multiplicities of the characteristic roots on the unit circle and conditional heteroscedasticity in the predictors and errors, are established. The simulation results corroborate the theoretical properties and show the superior performance of FHTD in model selection. FHTD is also applied to U.S. monthly housing starts and unemployment data to demonstrate its usefulness in practice.

EO311 Room 708 HIGH-DIMENSIONAL DATA ANALYSIS WITH CLUSTER STRUCTURE Chair: Antonio Elias

E0412: **Dynamic factor models with cluster structure**
Presenter: **Angela Caro Navarro**, Universidad Carlos III de Madrid, Spain
Co-authors: Maximo Camacho, Daniel Pena

Understanding whether the drivers of industrial energy prices are worldwide, group-specific or country-specific is a key issue in economics. This requires flexible econometric models to examine large data sets containing a significant variety of industrial sectors in different countries. To this end, an extension of a dynamic factor model with group structure is proposed to account for observable country-specific explanatory variables, and Monte Carlo simulations are developed to show its good finite sample performance. Using data from 12 industrial sectors in 30 countries during the period from 1995 to 2015, three drivers of energy prices are found: (i) a common factor, the main driving force, captures the worldwide dynamics; (ii) country-specific variables, mainly related to inflation and the use of renewable and waste resources; and (iii) group-specific factors, which are more related to country affiliation than to sector classification.

E0467: **Clustering and forecasting multiple functional time series**
Presenter: **Chen Tang**, The Australian National University, Australia
Co-authors: Han Lin Shang, Yanrong Yang

Modelling and forecasting homogeneous age-specific mortality rates of multiple countries could lead to improvements in long-term forecasting. Data fed into joint models are often grouped according to nominal attributes, which may still contain heterogeneity and deteriorate the forecast results. To address this issue, a novel clustering technique is proposed to pursue homogeneity among multiple functional time series based on functional panel data modelling. Common functional time series features can be extracted using a functional panel data model with fixed effects. These common features could be decomposed into two components: the functional time trend and the mode of variations of functions. The functional time trend reflects the dynamics across time, while the functional pattern captures the fluctuations within curves. The proposed clustering method searches for homogeneous age-specific mortality rates of multiple countries by accounting for both features. The proposed clustering technique outperforms other existing methods through a Monte Carlo simulation and could handle complicated cases with slow decaying eigenvalues. In empirical data analysis, it is found that the clustering results of age-specific mortality rates can be explained by the combination of geographic region, ethnic groups, and socioeconomic status. Further, it is shown that the model produces more accurate forecasts than several benchmark methods in forecasting age-specific mortality rates.

E0554: **Clustering multivariate functional data: An application of the multivariate epigraph and hypograph indexes**
Presenter: **Belen Pulido Bravo**, Universidad Carlos III de Madrid, Spain
Co-authors: Rosa Lillo, Alba Franco-Pereira

Dealing with functional data is a challenging problem since it involves working with an infinite-dimension problem. Multivariate functional data adds another layer of complexity by introducing multiple functions that may be related to each other. A major challenge when working with this type of data is how to order it, as there are potentially infinite observations to compare and contrast. There are several proposals in the literature. It is proposed to order multivariate functional data by generalizing the concepts of epigraph and hypograph indexes, which are very useful in the univariate functional data context, is proposed into the multivariate one. After that, clustering multivariate functional data by applying these indexes is proposed to reduce the dimension of the initial dataset. By applying clustering techniques to this final dataset, groups of functions that share similar features can be identified, and insights are gained into the underlying structure of the data. This approach has broad applications in fields such as neuroscience, economics, and environmental sciences, where multivariate functional data are common but challenging to analyze. This new approach has been evaluated against different options available in the literature and applied the methodology to a real data set.

E0861: **Quantile based functional principal component analysis**
Presenter: **Alvaro Mendez Civieta**, Universidad Carlos III de Madrid, Spain
Co-authors: Jeff Goldsmith, Ying Wei

The majority of methodologies in Functional Data Analysis FDA treat the curves as smooth functions observed with error and are centred on modelling the expected curves. However, this approach does not consider the within-subject variability and correlation of the data, as an expected value cannot reflect those. The method presented addresses this gap and seeks to capture the within-subject variability by modelling the subject-specific conditional quantiles. This objective assumes treating each subject as a single realization from its own underlying distribution. When this distribution is symmetric and no outliers in the data, the expected value provides very good results. However, in many applications, the distribution is skewed, with changes over the functional domain that are not reflected in the expected value but can be reflected by modelling the conditional quantiles. The functional quantile principal component analysis, FQPCA, is introduced, a dimensionality reduction technique that extends the concept of Functional Principal Components to the quantile regression framework, obtaining a model that can explain the subject-specific quantiles conditional on a set of loading functions. FQPCA can capture shifts in the scale and distribution of the data that may affect the quantiles but may not affect the mean and is also a robust methodology suitable for dealing with outliers, heteroscedastic data, or skewed data.

EO087 Room 709 RECENT DEVELOPMENTS IN INSURANCE STATISTICS Chair: Sangyeol Lee

E0946: **What is the average surplus before ruin?**
Presenter: **Jae-Kyung Woo**, UNSW Sydney, Australia
Co-authors: Eric Chi Kin Cheung, Runhuan Feng, Haibo Liu

The aim is to study the moments of the average surplus before ruin in a renewal risk process with a general interclaim time distribution. The average surplus before ruin is calculated as the area under the sample path divided by the ruin time, which provides a new ruin quantity of interest. However, the traditional approach of conditioning on the first claim event is no longer feasible because the ruin time appears in the denominator. To circumvent this, it is shown that the moments of the average surplus can be obtained by integrating the discounted moments of the area under the sample with respect to the force of interest. These discounted moments can then be determined using a moment-based discounted density similar to the one in Cheung (2013). Explicit formulas are also provided for the case where the claim amounts are a combination of exponentials.

E0961: **Tractable Poisson time-series models for experience rating**
Presenter: **Jae Youn Ahn**, Ewha Womans University, Korea, South

A new time series model for count data based on the observation-driven state space model is proposed. The proposed model has the advantage of being able to perform simple predictive mean and closed-form expression for the likelihood function and that the stationarity of the variance is guaranteed, whereas the existing observation-driven Bayesian state-space model does. Simulation study and real data analysis are accompanied to show the effectiveness of the proposed method.

E1002: **Attention is not not not explanation: With a focus on insurance ratemaking**
Presenter: **Kyungbae Park**, Kangwon National University, Korea, South
Co-authors: Jae Youn Ahn, Rosy Oh, Yang Lu, Dan Zhu

Attention mechanisms became a standard tool in NLP systems. In the classical attention mechanisms, attention weights served as the weight of input units; hence it is claimed that attention mechanisms provide interpretability. However, simulation studies in the series of recent studies arguably show that the explainability of the attention mechanism is questionable. The interpretability of the attention mechanism under the setting of insurance rate-making is investigated. Specifically, first, a mathematical argument is provided showing that the attention mechanism fails to provide explainability. Then, an alternative attention mechanism is provided where the explainability of the attention layer is guaranteed. A simulation study is accompanied to show the performance of the proposed method.

E1058: **A multivariate compound dynamic contagion process for infectious events**
Presenter: **Rosy Oh**, Korea Military Academy, Korea, South

The dynamic contagion process (DCP) is a generalization of the externally exciting Cox process with shot noise intensity and the self-exciting Hawkes process. To date, however, no theoretical work has been done on a multivariate dynamic contagion process (MDCP) and its compound process. The theoretical distributional properties for these processes are analyzed systematically, based on the piecewise deterministic Markov process theory. Exact simulation algorithms are provided for these processes. Numerical results show that these processes can be used for the modelling of aggregate loss and the number of infections arising from contagious catastrophic events.

EC292 Room 03 BIOSTATISTICS Chair: Michelle Miranda

E1112: **Linear biomarker combination for constrained classification**
Presenter: **Yijian Huang**, Emory University, United States

Multiple biomarkers are often combined to improve disease diagnosis. Unfortunately, the uniformly optimal combination, i.e., with respect to all reasonable performance metrics, requires excessive distributional modelling, to which the estimation can be sensitive. An alternative strategy is rather to pursue local optimality concerning a specific performance metric. Nevertheless, existing methods may not target the clinical utility of the intended medical test, which usually needs to operate above a certain sensitivity or specificity level, or do not have their statistical properties well studied and understood. The developments and investigation of a linear combination method will be discussed to maximize the clinical utility empirically for such a constrained classification. The combination coefficient is shown to have cube root asymptotics. The convergence rate and limiting distribution of the predictive performance are subsequently established, exhibiting the robustness of the method in comparison with others. An algorithm with sound statistical justification is devised for efficient and high-quality computation. Simulations corroborate the theoretical results and demonstrate good statistical and computational performance. An illustration with a clinical study on aggressive prostate cancer detection is provided.

E1131: **Bayesian profile regression for high-dimensional data: An application to osteoarthritis proteomic data**
Presenter: **Laura Bondi**, University of Cambridge and Bocconi University, United Kingdom
Co-authors: Brian Tom, Sylvia Richardson

There is a huge unmet need in osteoarthritis (OA), with an estimated 8.5 million people affected in the UK. It is regarded as a highly heterogeneous disease and is purported to exist in different forms. As part of the STEpUP OA collaboration, an academic-industry partnership, the work explores the molecular pathways of OA and aims at identifying subpopulations of patients homogeneous for protein marker profiles such that each cluster has a clinical meaning (outcome-guided clustering). Bayesian profile regression (model-based outcome-guided clustering approach) is carried out to identify clusters of protein marker profiles that are associated with clinically relevant outcomes, such as disease radiographic grade (low vs advanced). This clustering methodology can handle possibly inter-related explanatory variables and uses the information in both these explanatory variables (i.e. 6000 synovial protein markers) and the outcome to produce model-based clustering structures, where the uncertainty associated with these clustering structures and the number of clusters is reflected. Given the high dimensionality of the protein space, computational challenges arise when scaling profile regression in this context. The focus is on strategies for dimensionality reduction and variable selection, taking into account biological knowledge. Moreover, the influence of the clinical outcome to drive the clustering structure is investigated.

E1233: **A statistical framework for fine-mapping by leveraging genetic diversity and accounting for confounding bias**
Presenter: **Mingxuan Cai**, City University of Hong Kong, Hong Kong

Fine mapping prioritizes risk variants identified by genome-wide association studies (GWASs), serving as a critical step to uncover biological mechanisms underlying complex traits. However, several major challenges still remain for existing fine-mapping methods. First, the strong linkage disequilibrium among variants can limit fine-mapping's statistical power and resolution. Second, it is computationally expensive to search for multiple causal variants simultaneously. Third, the confounding bias hidden in GWAS summary statistics can produce spurious signals. To address these challenges, a statistical method is developed for cross-population fine-mapping (XMAP) by leveraging genetic diversity and accounting for confounding bias. Using cross-population GWAS summary statistics from global biobanks and genomic consortia shows that XMAP can achieve greater statistical power, better control of false positive rate, and substantially higher computational efficiency for identifying multiple causal signals compared to existing methods. Importantly, it is shown that the output of XMAP can be integrated with single-cell datasets, greatly improving the interpretation of putative causal variants in their cellular context at single-cell resolution.

E1166: **Biostatical models to investigate physiological mediators regarding cognition on cardio-metabolic risk factors**
Presenter: **Sujin Kang**, Imperial College London, United Kingdom

While there is mounting evidence of an association between cardiovascular disease and brain health, the evidence from longitudinal cohort studies is currently inconclusive, in part because of a lack of studies with appropriate physiological measures. We aimed to investigate whether the targeted urinary metabolites and brain. MRI-derived parameters are prognostic markers regarding cognition on cardio-metabolic risk factors. Within the CARDIA Study, targeted metabolomics data was analysed by NMR spectroscopy and LC-MS, MRI data and cognitive scores for up to 606 participants aged 48-60 years at Year 30 (2015-16) and Year 25 (2010-11). A path analysis was conducted to investigate mediators of adjusted cognition on a cardio-metabolic factor measured at Year 30. Bayesian multilevel models were employed to examine cognitive decline measured at Year 25 and Year 30. Phenylalanine, aminoadipic acid, and tryptophan were mediators showing positive associations on fasting glucose in terms of hyperglycaemia, while indole-3-acetic acid showed a negative association. Cerebral blood flow in the temporal lobe white matter left and left entorhinal area showed negative associations on waist circumference in terms of (abdominal) obesity in the cohort. The multiple bio-statistical analysis approach applied here with a follow-up study may be able to recognise complex multivariate epidemiologic, pathologic, and phenotypic relationships across the disease networks that are yet unidentified.

EC276 Room 04 SURVIVAL ANALYSIS Chair: Yoann Potiron

E0365: **Comparing aging patterns of k-out-of-n systems using second-order stochastic dominance**
Presenter: **Idir Arab**, Univeristy of Coimbra, Portugal
Co-authors: Tommaso Lando, Paulo Oliveira

The problem of comparing ageing patterns of k-out-of-n systems with i.i.d. components are explored using second-order stochastic dominance. The focus is on determining a stochastic ordering relationship between different order statistics and characterizing such relationships with respect to relative convexity. A hierarchy of reference functions is introduced that encompasses popular families of distributions, including increasing failure rate distributions. Sufficient dominance conditions are derived based on the identification of the class that includes the component lifetimes. The applicability of this method is discussed.

E1152: **Analysis of errors-in-variables competing risks data in discrete time**
Presenter: **Chi-Chung Wen**, Tamkang University, Taiwan

Analysis of competing risk data has been an important topic in survival analysis due to the need to account for the dependence among competing events. Also, event times are often recorded on discrete time scales, rendering the models tailored for discrete-time nature useful in the practice of survival analysis. Regression analysis with discrete-time competing risks data and the errors-in-variables issue where the covariates are prone to measurement errors are considered. Without assuming a distribution for the true covariate, we develop robust sufficient score methods for the cause-specific and subdistribution hazards models. TEfficient computation algorithms can implement the proposed estimators, and the associated large sample theories can be simply obtained.

E1292: **Estimating time-varying treatment effects on restricted mean survival time in large patient databases**
Presenter: **Chi Hyun Lee**, University of Massachusetts Amherst, United States

The restricted mean survival time (RMST), defined as the life expectancy up to a specific point, has recently attracted substantial attention as an alternative to the hazard ratio for quantifying the treatment effect in clinical studies. A flexible model is proposed to estimate the effect of treatment based on RMST. The effect of treatment is expressed as a function of restriction time better to characterize the dynamic trend of its effect on survivall. To account for possible heterogeneity across patients in large databases, the propensity scores are incorporated for receiving treatment into the model. Further, an ensemble approach is introduced to aggregate estimators constructed based on subsamples of the observed failure times. The finite sample performance of the proposed single model and ensemble-based approaches through simulations are evaluated, and the proposed methods are applied to the study of primary inflammatory breast cancer for assessing the effect of trimodality therapy on survival.

E1200: **Bayesian joint modelling of longitudinal and competing risks data with cause dependent masking**
Presenter: **Mahaveer Singh Panwar**, Banaras Hindu University, Varanasi, India

Joint modelling of longitudinal measurements and time-to-event data has received considerable attention, especially in the field of public health studies. The longitudinal and competing risks event processes, with a linear mixed effect model and cause-specific hazard model, are jointly analyzed, respectively. The two processes are associated with the shared random effect approach. In the competing risks data, the event's cause is not always known, leading to incomplete data concerning the cause. The cause of events for such individuals is said to be masked. It is assumed that the masking is not independent of the causes, and hence, our proposed joint model also deals with cause-dependent masking situations in competing risk data. The estimation of model parameters is carried out under the Bayesian paradigm as it is computationally more flexible against the model complexity. An extensive numerical study is performed to evaluate the efficacy of estimators obtained under the joint model. The Prostate, Lung, Colorectal, and Ovarian (PLCO) cancer prevention trial dataset illustrates the established methodology subsuming the dependent masking in competing risks data.

EC272 Room 102 MULTIVARIATE STATISTICS Chair: Sara Lopez Pintado

E0202: **Random-covariate-dependent rectangular reference regions under multivariate normality**
Presenter: **Raden Gerald Agustin**, University of the Philippines Cebu, Philippines
Co-authors: Michael Daniel Lucagbo

Reference intervals are among the most widely used decision-making tools in the medical field and are invaluable in interpreting laboratory test results. These intervals may depend on covariates such as age and sex. The covariate values are often random quantities since they are typically not controlled in reference interval determination studies. When several biochemical analytes are needed to diagnose the same condition, the use of combined univariate reference intervals is not recommendable since the analytes could be correlated. Instead, a multivariate reference region is necessary to consider the correlations among the analytes. Traditionally, multivariate reference regions have been constructed as ellipsoidal. However, such regions cannot detect the outlying ness of a specific analyte. Procedures are proposed to construct rectangular multivariate reference regions incorporating random covariate information from the subjects. The reference regions are computed in a multivariate normal setting, and the prediction region criterion is used. A parametric bootstrap approach is employed to calculate the required prediction factor. Numerical results show that the parametric bootstrap approach is entirely accurate, with coverage probabilities very close to the desired nominal value. Finally, the proposed method is applied to real-life data from a study to compute covariate-dependent reference regions for insulin-like growth factors.

E1154: **The influence function of scatter halfspace depth**
Presenter: **Germain Van Bever**, Universite de Namur, Belgium
Co-authors: Gaetan Louvet

Statistical depth provides robust nonparametric tools to analyze distributions. Depth functions indeed measure the adequacy of distributional parameters to underlying probability measures. In the location case, the celebrated (Tukey) halfspace depth has been widely studied, and its robustness properties are amply discussed. Recently, depth notions for scatter parameters have been defined and studied. The robustness properties of this latter depth function remain, however, largely unknown. The influence function of scatter halfspace depth is derived. Expressions are given in the known and unknown location cases under mild distributional assumptions. In the latter case, the expression allows disentangling the unknown location effect from the scatter contamination. The corresponding asymptotic variance is also provided.

E1243: **Spacing test for generalized lasso with full row rank of D: Fused lasso and trend filtering**
Presenter: **Rieko Tasaka**, Osaka University Graduate School, Japan
Co-authors: Ryosuke Shimmura, Joe Suzuki

A generalized lasso is considered. In generalized lasso, fused lasso and trend filtering are exceptional cases in that the matrix D in the objective function is full row rank. Fused lasso and trend filtering are methods for smoothing one-time and second-order differences. Given an N-dimensional observation vector y and a constant > 0, an N-dimensional vector is obtained that minimizes the least-squares regression equation multiplied by a penalty term. A novel approach is proposed to solve the problem from the post-selective inference (PSI) perspective. In particular, a PSI method is considered called the spacing test, considered previously for linear regression Lasso. The spacing test assumes LARS, which is an approximation of lasso. Spaing tests the set of selected variables in which the null hypothesis is that all the necessary variables have been selected. The main contribution of the research is to modify the spacing test for choosing the appropriate value, which validates the choice of the fused lasso. The R program is made to execute the procedure to examine whether the proposed method works even for large N. The proposed method can also be considered similar to trend filtering, but unlike fused lasso, it cannot be solved by simply adding variables, as in LARS. Hence, it needs to be devised differently from a fused lasso.

E1249: **Probabilistic and distance-based approaches for computing multivariate highest-density regions**
Presenter: **Nina Deliu**, Sapienza University of Rome; University of Cambridge, Italy
Co-authors: Liseo Brunero

Many statistical problems require estimating a density function, say f, from data samples. Multivariate highest-density regions (HDRs) are considered - i.e., minimum volume sets containing a given probability - typically computed using a density quantile approach. Nevertheless, the density estimation task is far from trivial, especially over increased dimensions and when data are sparse and exhibit complex structures (e.g., multimodalities or particular dependencies). This challenge is addressed by exploring alternative approaches to build HDRs that overcome direct multivariate density estimation. First, the density quantile method - currently implementable based on a consistent density estimator - is generalized to neighbourhood measures, i.e., measures that preserve the order induced in the sample by f. Second, it is elaborated on, and several suitable probabilistic - and distance-based measures are evaluated, such as the k-neighbourhood Euclidean distance. Third, motivated by the ubiquitous role of copula modelling in modern statistics, its use in probabilistic-based measures is explored. By separately modelling marginals and their (potentially complex) dependence structure - that is the copula - the multivariate density estimation and better capture data specificities can both be relaxed. Finally, a comprehensive comparison among the introduced measures is provided, and their implications for computing HDRs in real-world problems are discussed.

EC283 Room 201 NON- AND SEMI-PARAMETRIC METHODS Chair: Yoichi Nishiyama

E0281: **Non-parametric inference of spatial birth-death-move processes**
Presenter: **Ronan Le Guevel**, University of Rennes, France
Co-authors: Frederic Lavancier

The analysis of the spatiotemporal dynamics of proteins involved in exocytosis in cells is an important biological challenge. A new spatial birth-death-move process, where the birth and death dynamics depend on the current spatial configuration of the population and where individuals can move with possible interactions during their lifetime, is introduced in order to model this complex mechanism. A non-parametric kernel method, with continuous time observations as well as discrete-time observations, is presented for the estimation of the jump intensity function, involving, in most cases, a choice of a distance between point configurations and a choice of a window which will also be discussed. The procedure is illustrated with simulations and results on the biological dataset.

E1145: **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 with respect to a distribution or a data set by a number between 0 and 1 while satisfying certain postulates regarding invariance, monotonicity, and convexity. It constitutes a contemporary domain of rapid development to meet growing demand in various areas of 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 so that a limited training sample may easily result in poor out-of-sample performance. A generalized halfspace depth concept relying on the recent advances in distributionally robust optimization is proposed, where every halfspace is examined using the respective worst-case distribution in the Wasserstein ball centred at the empirical law. This new depth can be seen as a smoothed and regularized classical halfspace depth, which is retrieved as the radius of the Wasserstein ball vanishes. It inherits the main properties of the latter and, additionally, 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 theories are developed. In particular, the upper-level sets and the median region are studied, including their breakdown properties.

E1195: **Variance properties of local polynomial density estimators at the boundary: Application to manipulation testing**

Presenter: **Yuta Okamoto**, Kyoto University, Japan

Co-authors: Shunsuke Imai

Variance properties of the local polynomial density estimator at the boundary points are investigated. Asymptotic and theoretical non-asymptotic analyses reveal that kernel selection is crucial in contrast to common sense in nonparametric statistics. In particular, it is shown that the estimation accuracy can be devastatingly poor with the most used kernel functions, including the uniform and triangular kernels. More seriously, this finding also applies to the performance of the manipulation testing based on the estimator, which is the most accepted method in empirical economics. It is shown that the test has only disappointing statistical power and often fails to detect the unbalance in a running variable. However, these problematic natures are drastically resolved just by changing the kernel function to a less-known one. Numerous simulations and empirical applications are provided to highlight the empirical relevance, and these results strongly support the use of a carefully chosen kernel function.

E1309: **Nonparametric estimation of non-crossing quantile regression process with deep ReQU neural networks**

Presenter: **Guohao Shen**, The Hong Kong Polytechnic University, Hong Kong

Co-authors: Yuling Jiao, Yuanyuan Lin, Joel Horowitz, Jian Huang

A penalized nonparametric approach is proposed to estimate the quantile regression process (QRP) in a nonseparable model using rectified quadratic unit (ReQU) activated deep neural networks and introduce a novel penalty function to enforce the non-crossing of quantile regression curves. The non-asymptotic excess risk bounds for the estimated QRP are established, and the mean integrated squared error for the estimated QRP under mild smoothness and regularity conditions are derived. A new error bound for approximating C^s smooth functions with $s > 0$ and their derivatives using ReQU-activated neural networks is also developed to establish these non-asymptotic risk and estimation error bounds. This is a new approximation result for ReQU networks and is of independent interest, and may be useful in other problems. The numerical experiments demonstrate that the proposed method is competitive with or outperforms two existing methods, including methods using reproducing kernels and random forests for nonparametric quantile regression.

EC284 Room 203 APPLIED ECONOMETRICS II Chair: Masayuki Hirukawa

E1315: **Exploring profitability changes in hospitality industry: An econometric and statistical perspective**

Presenter: **Katerina Pericleous**, Cyprus University of Technology, Cyprus

Co-authors: Petros Kosmas, Antonis Theocharous, Elias Ioakimoglou, Hristo Andreev

The aim is to analyse profitability indicators within the hospitality industry comprehensively. This will be achieved by utilising a well-established empirical framework by studying the capital profitability of the Republic of Cyprus. Furthermore, the variations in the rate of return on fixed capital and identifies the key variables influencing profitability will be examined. The data analysed is from 2005 to 2020. Employing rigorous statistical analysis and econometric modelling, it untangles the intricate relationship between profitability indicators and the multifaceted factors shaping the financial performance of the hospitality sector. The method also endeavours to discern distinct capital valorisation and accumulation phases, providing valuable insights into the fundamental dynamics at play. The findings reveal significant implications for industry stakeholders, policymakers, and researchers, as they offer actionable recommendations and a comprehensive understanding of the profitability dynamics in the Republic of Cyprus hospitality industry. Additionally, it sets a stepping stone for future endeavours and contributes to advancing econometrics and statistics in the field.

E0440: **Can we have the cake and eat it too? The case for the top-floor units as a status good and an investment**

Presenter: **Chi Ho Tang**, Hong Kong Shue Yan University, Hong Kong

Co-authors: Ka Yui Leung

Located at the top of condominium buildings, Top Floor Units (TFU) offer superior views and privacy through the accessible roofs. The purpose is to study this status good empirically. It is found that (1) TFU interacts with the macroeconomy differently from the ordinary units, (2) TFU should not be included in the portfolio once the liquidity factor is considered, (3) the holding period-annualized return tradeoff of TFU differ significantly from the ordinary units, suggesting alternative investment strategies being used in TFU, and (4) the liquidity of the TFU segment is less stable than the ordinary units, and may therefore discourage short-term speculators.

E1014: **Impact of various weather data on leisure sales**

Presenter: **Tomas Tichy**, VSB-TU Ostrava, Czech Republic

Previous research is continued, and several more weather data are analyzed, specifically their possible relation to demand shocks. For example, it is already well known that with increasing pressure on performance and widely available data on customer behaviour and supply chain process, the entities are trying to optimize (minimize) the level of bounded capital while assuring at least a minimal level of benchmark parameters. When the data is extended, the decision about their quality (frequency) starts to be crucial. So a good estimate of future values can substantially increase the efficiency of the overall process. In particular, the dependency of several leisure activities on various weather data is considered; the results are compared with simple linear methods and truly observed temperatures. The utilization of fuzzy logic and if-then rules combined with complex databases show significant improvement over standard approaches.

E1208: **Firm behaviour in the European carbon market: Latent profile analysis on network indicators of transactions**

Presenter: **Marie Raude**, Climate Economics Chair/EconomiX Universite Paris Nanterre, France

Latent profile analysis (LPA) is carried out to identify different behaviour of firms in the European Union Emissions Trading System (EU ETS). The goal is to identify profiles based on firm transaction activity and not considering other characteristics, thus without any a priori bias on the type of behaviour a firm may have. This is especially relevant in the current context of suspicion of speculation in the market. The aim is to contribute to the debate on the potential detrimental role played by certain actors. Firm transaction behaviour is summarised using centrality and popularity indicators stemming from network theory. Estimated indicators include degree, strength, betweenness and eigenvector centralities, harmonic closeness and PageRank. Based on these indexes, LPA is performed to investigate the existence of different firm profiles and identify them. Firm characteristics are then added as covariates to explore the determinants of belonging to a certain group. Visualisation of the transactions network, along with the estimation of indicators, reveals that the network is highly polarised around a handful of financial firms that play the role of intermediary for compliance firms. The LPA further leads to the identification of three relevant profiles, with notably a popular profile characterised by extreme centrality values. Firm regulatory status and sectoral activity seem to be playing a significant role in determining profile assignment.

EC318 Room 503 BAYESIAN MODELLING AND INFERENCE Chair: Yasuhiro Omori

E0317: A generalization of the Dirichlet-multinomial regression model for microbiome counts

Presenter: **Sonia Migliorati**, University of Milano Bicocca, Italy
Co-authors: Andrea Ongaro, Roberto Ascari

Multinomial regression is a widespread tool to model microbiome counts as a function of environmental and biological covariates, whereas the Dirichlet-multinomial model represents an enhancement to cope with overdispersion. Though, both models often show a poor fit to real data due to their rigid dependence structure, which rules out the possibility of modelling positive associations among bacterial taxa and poor parameterization. A new regression model based on a mixture of Dirichlet-multinomial distributions is proposed. The model is a compound multinomial model, which is obtained by considering a (conditional) multinomial response and assigning an extended flexible Dirichlet distribution to its parameters. This new model succeeds in clearly identifying and interpreting relationships between taxa counts and covariates, allowing for possible positive associations among taxa too. Moreover, the mixture structure of the model naturally enables the identification of clusters of bacterial genera sharing similar biota compositions, which, in turn, can be associated with enterotypes. The analysis of a human gut microbiome dataset confirms the better performance of the new model concerning competitors. The analysis has been performed via bayesian inference, resorting to the Hamiltonian Monte Carlo algorithm, with a spike and slab approach to variable selection.

E0611: Bayesian inference for differential item functioning detection in a multiple-group IRT tree model

Presenter: **Yu-Wei Chang**, National Chengchi University, Taiwan
Co-authors: Cheng-Xin Yang

Group differences have practical implications in analyzing data from achievement tests or questionnaires. For example, whether two persons from different demographic groups, such as gender or race, with the same shopping preferences have different shopping habits on one aspect helps store managers better design their displays. Shopping habits and shopping preferences can be measured by items and some latent factors in a questionnaire, and the different shopping habits observed on an item are called differential item functioning (DIF). In the current study, a model that accounts for between-group differences, DIF, latent factors, and missing item response data simultaneously is developed by expanding a one-group item response tree model into a multiple-group model. Different from most of the present DIF studies, where one has to iteratively select anchor items and detect DIF items, DIF detection and parameter estimation simultaneously are achieved by properly reparameterizing model parameters and applying some spike-and-slab priors in Bayesian estimation. Simulation studies are conducted to illustrate the validation of the proposed estimation procedure and the efficiency of DIF detection. The proposed method is further applied to a real dataset for illustration.

E1081: Closed form Bayesian inferences for binary logistic regression with applications to American voter turnout

Presenter: **Kevin Dayaratna**, The Heritage Foundation, United States
Co-authors: Jesse Crosson, Chandler Hubbard

Understanding the factors influencing voter turnout is a fundamentally important question in public policy and political science research. Bayesian logistic regression models are useful for incorporating individual-level heterogeneity to answer these and many other questions. When these questions involve incorporating individual-level heterogeneity for large data sets that include many demographic and ethnic subgroups, however, standard Markov Chain Monte Carlo (MCMC) sampling methods to estimate such models can be quite slow and impractical to perform in a reasonable amount of time. An innovative closed form Empirical Bayesian approach is presented that is significantly faster than MCMC methods, thus enabling the estimation of voter turnout models previously considered computationally infeasible. The results shed light on factors impacting voter turnout data in the 2000, 2004, and 2008 presidential elections. It is concluded with a discussion of these factors and the associated policy implications. It is emphasized, however, that although the application is to the social sciences, this approach is fully generalizable to the myriads of other fields involving statistical models with binary dependent variables and high-dimensional parameter spaces as well.

E1146: Beta four parameter generalized linear mixed model using a Bayesian approach to predict paddy productivity

Presenter: **Dian Kusumaningrum Hermanto**, Prasetiya Mulya University/IPB University, Indonesia

A new Generalized Linear Mixed Model (GLMM) for a response variable having a beta four-parameter distribution based on the Bayesian approach is introduced. The framework expanded the beta four-parameter regression model to incorporate random effects in the model. The methodology is illustrated through simulations and applied to predict paddy productivity. Paddy productivity ranges between a minimum and maximum value; therefore, assuming a beta four-parameter distribution is most appropriate. Farmer survey and Sentinel satellite imagery data were used as covariates. The response variable was based on plots surveyed by the Central Bureau of Statistics (CBS) in Central Kalimantan, Indonesia. Results showed that this approach could overcome complicated back-transform processes, difficulties in interpreting the results obtained, and bias in the estimated parameters if the transformation to a standard beta distribution process was to be applied. In predicting paddy productivity, results were proven to be more accurate. Thus, it will be a beneficiary as an early warning system for food insecurity, a reference for the food self-sufficiency program, and a basis to calculate premiums and risks for the alternative Area Yield Index (AYI) crop insurance policy.

EC279 Room 506 MACHINE LEARNING IN ECONOMICS AND FINANCE Chair: Jeffrey Bohn

E0740: Boosting time-series prediction performance for inflation indicators

Presenter: **Jeffrey Bohn**, UC Berkeley, United States

As a risk, trading, strategy, and decision-support systems have become more deeply integrated into financial services firms' workflows, predicting a collection of economic and market indicators becomes even more critical to support these systems than in the past. At the same time, the underlying processes that drive economies and markets have become increasingly dynamic, given they are more likely to be subject to rapid successions of regime changes. Conventional curve-fitting frameworks that assume linear/log-linear, stable relationships continue to exhibit degraded predictive performance. Fortunately, innovations are being found in machine-learning algorithmic frameworks that lead to a collection of promising techniques defined as boosted trees or gradient boosting. These boosting methods better capture underlying non-linear data relationships in ways that can materially improve predictive performance for economic and market indicators. Some of the newer boosting methods and report compelling results will be described for predicting inflation with a subset of these methods. An approach called entropy boosting will be introduced.

E1088: Corporate bond return prediction: An ensemble learning approach

Presenter: **Albert Zhao**, Nankai University, China
Co-authors: Shan Jiang

Using corporate bond, Treasury, and stock markets predictors, it is found that corporate bond returns are predictable based on an ensemble learning approach, Stacking. Combining various linear and nonlinear models, it is shown that Stacking generates more significant predictive power across bond ratings and maturities than most existing methods did in terms of statistical measure and economic gain. The most critical factors in corporate bond return prediction stem from Treasury, stock, and corporate bond markets jointly. While the overall performances of different Stacking models are satisfactory, simpler Stacking models perform best.

E0357: **Determinants of adoption of robo-advisory in banking services**
Presenter: **Witold Orzeszko**, Nicolaus Copernicus University in Torun, Poland
Co-authors: Dariusz Piotrowski

Robo-advisory is an example of the use of artificial intelligence technology in the area of finance. The current significance of robo-advisory to the financial sector is minor or marginal and boils down to formulating recommendations and implementing investment strategies. The ongoing digital transformation of the economy leads us to believe that this technology will be more widely used shortly with banking products. The aim is to identify factors significantly influencing bank customers' intention to use robo-advisory. Poland, a country where the banking services market is one of the largest and most developed in Central and Eastern Europe, is covered. Primary data was obtained through a survey conducted on a representative sample of 911 respondents aged 18-65. Using a multilevel ordered logit model and methods based on machine learning algorithms, the authors identified variables relating to the demographic and socio-economic characteristics, behaviours, and attitudes of consumers that primarily determine respondents' adoption of robo-advisory. The practical aspect of the work takes the form of recommendations formulated based on the results that can be used in the implementation of robo-advisory by the banking sector.

E1009: **Deep impulse control**
Presenter: **Bowen Jia**, The Chinese University of Hong Kong, Hong Kong
Co-authors: Hoi Ying Wong

A novel deep learning framework is developed to estimate the optimal control policy for the impulse control problem. Using deep neural networks, this numerical method allows for a general class of stochastic processes and consideration of uncontrollable co-integrated processes. The method also applies to high-dimensional cases for controllable and uncontrollable stochastic processes. This method can solve the optimal policy for a wide range of impulse control problems, such as irreversible reinsurance, interest rate intervention and stochastic inventory control.

EC295 Room 701 FINANCIAL ECONOMETRICS II Chair: Zudi Lu

E0214: **Testing beta constancy in capital asset pricing models**
Presenter: **Luis Antonio Arteaga Molina**, Universidad de Cantabria, Spain
Co-authors: Juan Manuel Rodriguez-Poo

A methodology for testing coefficient constancy in varying coefficient capital asset pricing models with endogenous regressors is proposed. The testing procedure is defined as a generalized likelihood ratio that compares the restricted and unrestricted sum of squared residuals. As a by-product, a nonparametric method that considers the endogenous nature of the regressors has developed to estimate the prices of risk; besides, the asymptotic properties of the estimators are established. The finite sample properties of the test by means of Monte Carlo experiments study and using critical values and p-values estimated using a bootstrap technique are also investigated. Finally, the test is applied to the Fama and French's model using a Fama-French 6 portfolio, sorted by size and book-to-market.

E1126: **Distributional asymmetries and currency returns**
Presenter: **Josef Kurka**, UTIA AV CR, v.v.i., Czech Republic

Numerous strands of literature have tried to explain the empirical failure of uncovered interest rate parity. The aim is to provide a risk-based explanation of this phenomenon that should vastly contribute to the currency pricing literature. The Extreme Volatility Risk Factor building on the crucial information for both stocks and currency pricing (e.g. the peso problems), is proposed that are contained in the tails of the distribution. Using a dataset of the 19 largest currencies, the time series of cross-sectional Average, Extreme Low and Extreme High Volatility is constructed. Preliminary empirical results uncover that especially Extreme High Volatility is priced in the cross-section of currency returns on top of Average Idiosyncratic Volatility.

E1229: **A study on asset price bubble dynamics: Explosive trend or quadratic variation?**
Presenter: **Simon Kwok**, University of Sydney, Australia
Co-authors: Robert Jarrow

The aim is to posit that when an asset exhibits a bubble, the time series of its prices can explode with positive probability if a quadratic variation (QV) risk premium is large enough. This QV channel for bubble explosion is new to the literature. Based on the local martingale theory of bubbles, sufficient conditions under which this QV explosion can occur are provided. Another possible explosion is also identified due to an autoregressive (AR) drift. Using the S&P 500 index and a sample of individual stocks over 1996-2021, the existence of price bubbles is documented and tested for price explosions. Almost all price explosion episodes discovered are associated with QV and not the AR drift channel.

E1215: **Measuring systemic risk with non-exchangeable dependence**
Presenter: **Andreas Heinen**, Universite de Cergy Pontoise, France
Co-authors: Sangwon Lee

Non-exchangeable dependence breaks the symmetry between the response of an individual firm to market distress and the market's reaction to individual firm distress. A non-exchangeable bivariate copula is used to model the joint distribution of the daily returns of a set of major U.S. financial institutions and of the market index for the 2001-2016 period. Based on this model, systemic risk measures such as CoVaR, Exposure-CoVaR, and MES are computed. More specifically, both static and dynamic versions of a non-exchangeable Clayton copula are estimated, combined with parametric and non-parametric marginal GARCH models for the returns. Further, the systemic risk measures, both in- and out-of-sample, are backtested. It is found that measuring systemic risk using a non-exchangeable copula outperforms its exchangeable counterpart.

EC296 Room 704 ECONOMETRIC AND STATISTICAL MODELLING Chair: Jouchi Nakajima

E1165: **Reexamination of bargaining power in the distribution channel under possible price pass-through behaviors of retailers**
Presenter: **Tomoki Matsumoto**, Nara Institute of Science and Technology, Japan
Co-authors: Tomohito Kamai, Yuichiro Kanazawa

The purpose is to investigate the determinants of relative power within a distribution channel by incorporating the price pass-through behaviour of a retailer into a Nash bargaining with a manufacturer and under the assumption of retail price observability for manufacturers. First, the retail margins are derived from a retailer maximizing its profit, assuming that the retailer anticipates the manufacturer's profit-maximizing response to the price pass-through behaviour. The manufacturer margins are then derived based on the generalized Nash bargaining with the retailer, where the parties negotiate the wholesale price. Finally, conditions are identified under which the bargaining power parameter is well-defined based on the values of retailer and manufacturer margins and the parameter describing the degree of price pass-through. A toy Bayesian estimation example using daily scanner panel data for canned tuna at a single retail chain in western Tokyo, Japan, demonstrates not only the applicability of the model but also a starkly contrasting result obtained without the price pass-through behaviour and the retail price observability for manufacturers.

E1036: **What drives cryptocurrency returns? A sparse statistical jump model approach**
Presenter: **Federico Cortese**, University of Milano-Bicocca, Italy
Co-authors: Erik Lindstrom, Petter Kolm

The statistical sparse jump model, a recently developed, robust and interpretable regime-switching model, is used to analyze the factors driving the return dynamics of the largest cryptocurrencies. This method simultaneously incorporates feature selection, parameter estimation, and state classification. A wide range of candidate features is considered, including cryptocurrency, sentiment, and financial market-based time series that are known to influence cryptocurrency returns. The empirical analysis demonstrates that a three-state model provides a good representation of the cryptocurrency return dynamics. The latent states are interpreted as a bull, neutral, and bear market regimes, respectively. Through the data-driven feature selection approach, the significant factors are identified, and insignificant ones are excluded. The results indicate that within the candidate features, the first moments of returns, features indicating trends and reversal signals, market activity, and public attention are key drivers of crypto market dynamics.

E1045: **Nowcasting GDP with factor-augmented high-dimensional MIDAS regression**
Presenter: **Jonas Striaukas**, Copenhagen Business School, Denmark

A factor-augmented high-dimensional mixed-frequency regression model is introduced to nowcast the US GDP growth. The new approach builds on the literature of nowcasting using sparse methods and goes beyond it by combining sparse regression with factor models. The estimator's convergence rates in a time series context are derived, allowing for mixing processes and heavier than exponential tails. The new technique is applied to nowcast the US GDP growth, and among other insights, it is found that it significantly improves over a range of more traditional nowcasting methods, which are based on either sparse regression or factor models, but not both.

E1329: **On the estimation of parameters of fractional Poisson processes**
Presenter: **Aditya Maheshwari**, Indian Institute of Management Indore, India

The Fractional Poisson Process (FPP) plays a pivotal role in modeling systems with long-range dependence. We will apply the Method of Moments for estimating the parameters that characterize the FPP, particularly emphasising the index of dispersion and the memory parameter. By generating sample paths through simulation, we will validate the effectiveness of the method in estimating these parameters. Through comparative analysis and simulations, we aim to demonstrate the robustness and practicality of this approach. The presentation aims to offer a methodological guide for researchers and practitioners to efficiently estimate FPP parameters, providing them with a powerful tool for understanding and modeling complex systems with long-range dependent properties.

EP327 Room Poster session III POSTER SESSION III Chair: Cristian Gatu

E1242: **Empirical analysis on characteristics of Japanese consumer behaviors based on the consumption values**
Presenter: **Zhejun Wang**, Doshisha University, Japan
Co-authors: Yuejun Zheng, Ryozo Yoshino

Most Japanese consumers give priority to the quality of goods or services, new goods, and fashion. The aim is to clarify the characteristics of Japanese consumer behaviour in connection with choosing durable, semi-durable, and nondurable goods, also its influential factors, using consumption values. Consumption values are used to explain the reasons for purchasing specific goods or brands, including functional, emotional, social, epistemic, and conditional values. To examine the selection criteria for purchasing behaviour for the above three categories of goods, the research data was collected from 1,080 Japanese between 18 and 75 years old, conducted by an online survey in February 2023. The results of data analysis have shown that Japanese consumers tend to prioritize quality over price when purchasing durable and semi-durable goods, while for nondurable goods, they tend to prioritize price. In addition, Japanese consumers focus on whether the components/materials are gentle to the health when they choose semi-durable and nondurable goods, but their emphasis is on energy-saving for durable goods. On the other hand, the common characteristics of the purchasing behaviour for the three categories of goods are personal preference over a trend, the preference for classic goods, and lower interest in diverse and novel functions.

E1246: **Utilizing latent space representation for clustering chronic kidney disease subtypes via electronic health records**
Presenter: **Ren-Hua Chung**, National Health Research Institutes, Taiwan
Co-authors: Djeane Debora Onthoni, Kuei-Yuan Lan, Tsung-Hsien Huang , Ying-Erh Chen

Chorionic Kidney Disease (CKD) is a globally prevalent, multifaceted disease, with its root causes varying among patients, complicating the analysis, treatment, and prognosis prediction. The Electronic Health Record (EHR), a valuable data comprising diverse and longitudinal medical data, were utilized to scrutinize CKD subtypes. Nevertheless, the EHR data's high dimensionality, heterogeneity, and incomplete time series posed challenges. Considering the EHR data's chronological nature, an end-to-end framework was devised to cluster CKD subtypes, considering the time gap between patient visits. The framework, implemented using UK Biobank's EHR dataset, encompasses three stages: data preprocessing, transformation, and clustering. The Convolutional Autoencoder (ConvAE) architecture is employed to convert preprocessed data into a low-dimensional format, which is subsequently clustered using Principal Component Analysis (PCA) and K-means algorithms. The efficacy of the transformation and clustering steps is evaluated through accuracy, Silhouette, Purity, and Entropy scores. High scores confirmed the framework's efficacy, enabling us to decipher clinical patterns for a nuanced understanding of each CKD subtype within the respective clusters.

E0311: **Simultaneous component decomposition and anomaly detection in financial time series**
Presenter: **Subin Jeong**, Chungnam National University, Korea, South
Co-authors: Minsu Park

Anomaly detection algorithms in financial time series have been developed through various studies. In time series data, not only seasonality and trend but also unknown fluctuations such as noise are included. An algorithm that can detect even skewed points is proposed with high frequandwell as removing components such as seasonal-trend in time series data. The proposed algorithm goes through two processes. First, a noise-robust signal decomposition method is applied using the statistical, empirical mode decomposition technique to decompose signals into intrinsic mode functions with unique frequencies, filtering out low-frequency signals such as seasonality and trend. Second, a generalized outlier detection approach that can be applied to skewed distributions was used through the first intrinsic mode function among the decomposed signals. Through various real data and simulated data, the proposed algorithm properly detects the influence values generated in the sparsely dense part of the asymmetric distribution, and the smoothing spline-based empirical mode decomposition method clearly decomposes the signals between high and low frequencies, resulting in a good performance. Through this method, the proposed algorithm is expected to be effectively applied to detect anomalies in nonlinear, non-stationary, and skewed time series data with trend and seasonal variations.

E0725: **Classifying Alzheimers Disease patients and identifying related BOLD signals using penalized logistic regression**
Presenter: **Hyeonjeong Lim**, Chungnam National University, Korea, South
Co-authors: Eunjee Lee, Jeong Yeon Park

Alzheimers Disease is one of the most prevalent types of dementia. As there is currently no complete cure for AD, it is crucial to detect and treat it early with proper care. Blood-Oxygen-Level Dependent (BOLD) signals can be used to identify abnormal patterns of brain activity in patients, which can facilitate early diagnosis of AD. Therefore, this study aims to explore the BOLD signals associated with the onset of AD and construct a model for classifying AD and NC patients based on these signals. We analyze the fMRI data provided by ADNI (Alzheimers Disease Neuroimaging Initiative). The study was conducted on 307 patients, each with 116 BOLD signals and corresponding demographic information. We extract the functional characteristics of the 116 BOLD signals by using functional principal component analysis (PCA) to calculate PC scores. We use the PC scores as explanatory variables in a logistic regression model. We consider LASSO, elastic-net, and SCAD penalties for variable selection. The prediction performance of the proposed method is compared with that of competing methods, including decision tree, random forest, and boosting models. We conduct a receiver operating characteristic (ROC) analysis to evaluate the model selection performance. As a result, we can identify BOLD signals related to Alzheimers Disease and proactively classify AD and NC by using the proposed model.

E0338: **Setting of optimal process conditions for a diaphragm rate of change using DOE**

Presenter: **Yong Hyun Um**, Chungnam National University, Korea, South

Co-authors: Min Koo Lee

The purpose is to find the optimal process conditions for the diaphragm rate of change using the design of experiments. Input variables are temperature, delay after pressure, degassing, and input amount, and output variables are outer diameter, inner diameter, depth, and thickness. A non-repetitive factorial design was used to investigate the effect of input variables on diaphragm dimensions. The regression equation obtained from the significant terms showed that the depth of the diaphragm decreased with higher temperature and more degassing. Response optimization was used to determine the optimal values for each factor to achieve the target dimension. The rate of change was calculated to mould the diaphragm into a desired dimension, and the obtained result is expected to have high reproducibility as it has a high R-square. This method provides a useful method for accurately manufacturing diaphragms using a new mould based on the obtained rate of change.

E0836: **Investigating resting-state fMRI for Alzheimer’s disease identification through functional data analysis**

Presenter: **Ido Ji**, Chungnam national university, Korea, South

Co-authors: Eunjee Lee

Alzheimer’s disease (AD) is a debilitating neurodegenerative disorder affecting millions worldwide. Early detection and accurate diagnosis of AD are crucial for effective intervention and disease management. The potential of functional data analysis (FDA) is investigated using blood oxygenation level-dependent (BOLD) signals from resting-state functional magnetic resonance imaging (rs-fMRI) for the classification of Alzheimer’s disease patients and healthy controls, as well as exploring brain regions associated with AD progression. Since FDA provides a powerful framework for analyzing complex biological signals, the methods were applied to extract relevant features from rs-fMRI and employed for classification purposes. The results demonstrate the effectiveness of the FDA in AD research using rs-fMRI data and, more importantly, reveal brain regions that may play a significant role in AD progression. This discovery could shed light on new neural mechanisms underlying AD and has potential implications for early diagnosis and targeted interventions.

E1330: **Prediction of PM10 in Seoul, Korea using Bayesian networks**

Presenter: **Man-Suk Oh**, Ewha Womans University, Korea, South

Recent studies revealed that fine dust in ambient air might cause various health problems, such as respiratory diseases and cancer. To prevent the toxic effects of fine dust, it is important to predict the concentration of fine dust in advance and to identify factors that are closely related to fine dust. We developed a Bayesian network model for predicting PM10 concentration in Seoul, Korea, and visualized the relationship between important factors. The network was trained by using air quality and meteorological data collected in Seoul between 2018 and 2021. The results showed that current PM10 concentration, season, and carbon monoxide (CO) were the top 3 effective factors in predicting PM10 concentration in 24 hours in Seoul and that there were interactive effects.

Authors Index