CFE-CMStatistics 2022

Looking forward, the CFE-CMStatistics 2022 will take place at King's College London, UK, 17-19 December 2022. Tutorials will be organized the 16th of December 2022. You are invited to actively participate in these events. Propositions for organizing tracks and sessions are strongly encouraged. For further information please contact info@cmstatistics.org.

16th International Conference on Computational and Financial Econometrics (CFE 2022)
http://www.CFEnetwork.org/CFE2022

This conference invites presentations that contain either computational or financial econometrics, or empirical finance components. Computational and financial econometrics have been of interest for many researchers in economics, finance, statistics, mathematics and computing. Financial time series analysis focusses on efficient and robust portfolio allocation over time, asset valuations with emphasis on option pricing, volatility measurement, modelling market microstructure effects and credit risk. Apart from theoretical developments, financial time series analysis also has a high empirical content measuring risk and return. The computational aspects of such analysis are of crucial importance since one typically deals with high-dimensional problems and a large number of observations. Existing algorithms often do not utilize the best computational techniques for efficiency, stability, or conditioning. Furthermore, environments for conducting econometrics are inherently computer-based. Integrated econometrics packages have grown well over the years, but still have much room for development.

15th International Conference of the ERCIM Working Group on Computational and Methodological Statistics (CMStatistics 2022)
http://www.CMStatistics.org/CMStatistics2022

This conference invites presentations focussed on all aspects of statistics. Of particular interest is research in important areas of statistical applications where both computing techniques and numerical methods have a major impact. All aspects of statistics which make use, directly or indirectly, of computing will be considered, as well as applications of statistics in diverse disciplines (e.g. economics, medicine and epidemiology, biology, finance, physics, chemistry, climatology and communication).

Publications

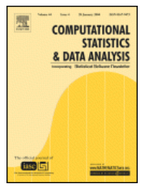
Econometrics and Statistics (EcoSta), published by Elsevier (http://www.elsevier.com/locate/ecosta), is the official journal of the Computational and Financial Econometrics and Computational and Methodological Statistics networks. It publishes research papers in all aspects of econometrics and statistics, and comprises two sections, namely,Part A: Econometrics Part B: Statistics. The EcoSta journal publishes as a supplement the Annals of Computational and Financial Econometrics.

Computational Statistics & Data Analysis (http://www.elsevier.com/locate/csda) is an official journal of CMStatistics. It publishes special issues and the Annals of Statistical Data Science in collaboration with expert teams drawn from the network Computational and Methodological Statistics. The special issues and Annals of Statistical Data Science provide an outlet for publications of high quality that are expected to have a significant impact in the subject areas of computing, methodological statistics, data analysis, and applied statistics.

Papers containing significant novel components in econometrics or statistics are encouraged to be submitted for publication in special or regular peer-reviewed issues of the new Elsevier journal Econometrics and Statistics. Papers containing strong computational statistics, or substantive data-analytic elements can also be submitted to special or regular peer-reviewed issues of the journal Computational Statistics & Data Analysis (CSDA).

CFE-CMStatistics 2022, 17-19 December 2022

King's College London, UK



Book of Abstracts

CFE-CMStatistics 2021

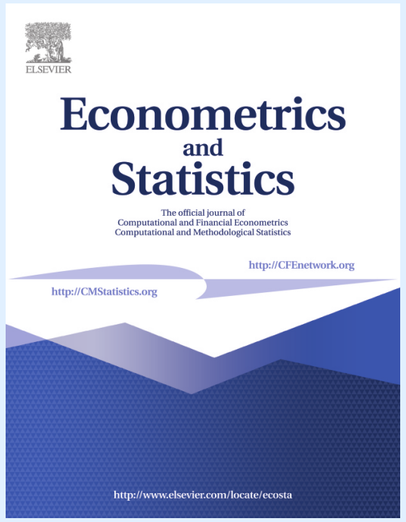
15th International Conference on Computational and Financial Econometrics

14th International Conference of the ERCIM Working Group on Computational and Methodological Statistics

International networks of
Computational and Financial Econometrics, CFEnetwork
Computational and Methodological Statistics, CMStatistics

18-20 December 2021

King's College London, UK



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

Part A: Econometrics. Emphasis is given to methodological and theoretical papers containing substantial derivations in econometrics or showing the potential to significantly impact the broad area of econometrics. Topics of interest include the estimation of econometric models and associated inference, model selection, panel data, measurement error, Bayesian methods, and time series analyses. Simulations are considered when they involve an original methodology. Innovative papers in financial econometrics and its applications are considered. The covered topics include portfolio allocation, option pricing, quantitative risk management, systemic risk and market microstructure. Interest is also given to well-founded applied econometric studies that demonstrate the practicality of new procedures and models. Such studies should involve the rigorous application of statistical techniques, including estimation, inference and forecasting. Topics include volatility and risk, credit risk, pricing models, portfolio management, and emerging markets. Innovative contributions in empirical finance and financial data analysis that use advanced statistical methods are encouraged. The results of the submissions should be replicable. Applications consisting only of routine calculations are not of interest to the journal.

Part B: Statistics. Papers providing important original contributions to methodological statistics inspired by applications are considered for this section. Papers dealing, directly or indirectly, with computational and technical elements are particularly encouraged. These cover developments concerning issues of, e.g., high-dimensionality, re-sampling, dependence, robustness or filtering. In general, the interaction of mathematical methods and numerical implementations for the analysis of large and/or complex datasets arising in areas such as medicine, epidemiology, biology, psychology, climatology and communication is considered. 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 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. The journal publishes as a supplement the Annals of Computational and Financial Econometrics.

PROGRAMME AND ABSTRACTS

15th International Conference on

Computational and Financial Econometrics (CFE 2021)

http://www.cfenetwork.org/CFE2021

and

14th International Conference of the
ERCIM (European Research Consortium for Informatics and Mathematics) Working Group on

Computational and Methodological Statistics (CMStatistics 2021)

http://www.cmstatistics.org/CMStatistics2021

King's College London, UK

18 – 20 December 2021

ISBN 978-9925-7812-5-6

International Organizing Committee:
Ana Colubi, Erricos Kontoghiorghes and Manfred Deistler.

CFE 2021 Co-chairs:
Degui Li, Alessia Paccagnini, Mark Podolskij and Martin Wagner.

CFE 2021 Programme Committee:

Cristina Amado, David Ardia, Juan Arismendi-Zambrano, Josu Arteche, Monica Billio, Kris Boudt, Scott Brave, Ruijun Bu, Andrew Butters, Massimiliano Caporin, Jia Chen, Julien Chevallier, Christian Conrad, Christa Cuchiero, Serge Darolles, Luca De Angelis, Ying Fang, Catherine Forbes, Markus Fritsch, Ana-Maria Fuertes, Guidolin Guidolin, Daniel J. Henderson, Douglas Hodgson, Rustam Ibragimov, Laura Jackson Young, Maria Kalli, Ekaterina Kazak, Edward Knotek, Anders Kock, Robinson Kruse-Becher, Robert Kunst, Emese Lazar, Leonida Leonida, Yifan Li, Marco Lorusso, Richard Luger, Ingmar Nolte, Michael T Owyang, Christopher Parmeter, Sandra Paterlini, Manuela Pedio, Peter Pedroni, Indeewara BS Perera, Tommaso Proietti, Artem Prokhorov, Francesco Ravazzolo, James Reade, Julia Schaumburg, Joachim Schnurbus, Willi Semmler, Etsuro Shioji, Michael Smith, Mike So, Leopold Soegner, Marco Maria Sorge, Robert Taylor, Helena Veiga and Toshiaki Watanabe.

CMStatistics 2021 Co-chairs:
Xavier de Luna, Kalliopi Mylona, Marianna Pensky and Javier Rubio.

CMStatistics 2021 Programme Committee:

Joseph Antonelli, Alexander Aue, Soutir Bandyopadhyay, Andres F. Barrientos, Anastassia Baxevani, Rebecca Betensky, Michela Bia, Markus Bibinger, Natalia Bochkina, Scott Bruce, Cristina Butucea, Antonio Canale, Eva Cantoni, Jeng-Min Chiou, Eliana Christou, Sophie Dabo, Mike Daniels, Rob Deardon, Walter Dempsey, Reza Drikvandi, Yang Feng, Maria Brigida Ferraro, Robert Gaunt, Christian Genest, Steven Gilmour, Virgilio Gomez Rubio, Michele Guindani, Sebastien Haneuse, Siegfried Hoermann, Karel Hron, Piotr Jaworski, Binyan Jiang, M. Dolores Jimenez-Gamero, Jiashun Jin, John Kornak, Marie Kratz, Christophe Ley, Shujie Ma, Saumen Mandal, Hiroki Masuda, Daniel J. McDonald, Geoff McLachlan, Ramses H. Mena, Cristina Mollica, Erica Moodie, Gourab Mukherjee, Farouk Nathoo, Daniel Nevo, Klaus Nordhausen, Alex Petersen, Michael Pitt, Wolfgang Polonik, Xinghao Qiao, Matias Quiroz, Monia Ranalli, Marialuisa Restaino, Jonathan Schildcrout, Mireille Schnitzer, Jian Qing Shi, Elena Stanghellini, Gilles Stupfler, James Taylor, Ingeborg Waernbaum, Helga Wagner, Nakahiro Yoshida, Anderson Ye Zhang, Yichuan Zhao, Yeying Zhu and Anne van Delft.

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

Dear Friends and Colleagues,

We are delighted to have the opportunity to meet in these difficult times. We are still passing through extraordinary events that are significantly affecting our personal and professional lives. In order to cope with the uncertainty caused by the pandemic, we have implemented a hybrid format, so that the participants can select to participate in-person or virtually according to their circumstances. Some sessions were planned to be virtually from the beginning, but most were planned to be hybrid or in-person. Unfortunately, the last weeks have forced many participants to cancel their travel plans and opt for the virtual mode. The programme has been dynamically adapted to that with the hope of allowing the participants to present their results and network in the best possible way. Despite the organization challenges, we are happy to welcome the over 1650 participants warmly.

The 15th International Conference on *Computational and Financial Econometrics* (CFE 2021) and the 14th International Conference of the ERCIM Working Group on *Computational and Methodological Statistics* (CMStatistics 2021) have shown, once again, the relevance of the CFE-CMStatistics meetings at the interface of statistics, econometrics, empirical finance and computing. The conference aims at bringing together researchers and practitioners to discuss recent developments in computational methods for economics, finance, and statistics. The CFE-CMStatistics 2021 programme consists of near 400 sessions, five plenary talks and about 1600 presentations.

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

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

The Elsevier journal Econometrics and Statistics (EcoSta) was inaugurated in 2017. The EcoSta is the official journal of the networks of Computational and Financial Econometrics (CFEnetwork) and of Computational and Methodological Statistics (CMStatistics). It publishes research papers in all aspects of econometrics and statistics, and it comprises two sections, namely, Part A: Econometrics and Part B: Statistics. The participants are encouraged to submit their papers to special or regular peer-reviewed issues of EcoSta and its supplement Annals of Computational and Financial Econometrics.

The CMStatistics has also commenced The Annals of Statistical Data Science (SDS), which will be published as a supplement of the Elsevier journal Computational Statistics & Data Analysis (CSDA). The CSDA is also the official journal of CMStatistics. You are encouraged to submit your papers to the Annals of Statistical Data Science or regular peer-reviewed issues of CSDA.

Looking forward, the CFE-CMStatistics 2022 will be held at King's College London, from Saturday the 17th of December 2022 to Monday the 19th of December 2022. Tutorials will take place on Friday the 16th of December 2022. You are invited and encouraged to participate in these events actively.

We wish you a productive and stimulating conference.

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

CMStatistics: ERCIM Working Group on COMPUTATIONAL AND METHODOLOGICAL STATISTICS

http://www.cmstatistics.org

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

Specialized teams

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

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

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

CFEnetwork COMPUTATIONAL AND FINANCIAL ECONOMETRICS

http://www.CFEnetwork.org

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

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

SCHEDULE (GMT)

2021-12-18

A - Opening and Keynote
CFE - CMStatistics
08:15 - 09:15

B
CFE - CMStatistics
09:25 - 10:40

Coffee Break
10:40 - 11:10

C
CFE - CMStatistics
11:10 - 12:50

Lunch Break
12:50 - 14:20

D
CFE - CMStatistics
14:20 - 16:00

Coffee Break
16:00 - 16:30

E
CFE - CMStatistics
16:30 - 18:35

F
CFE - CMStatistics
18:45 - 20:00

Welcome Reception
20:00 - 21:30

2021-12-19

G
CFE - CMStatistics
08:15 - 09:55

Coffee Break
09:55 - 10:25

H
CFE - CMStatistics
10:25 - 12:30

Lunch Break
12:30 - 14:00

I
CFE - CMStatistics
14:00 - 15:40

Coffee Break
15:40 - 16:10

J
CFE - CMStatistics
16:10 - 17:25

K
CFE - CMStatistics
17:35 - 19:15

L - Keynote
CFE - CMStatistics
19:25 - 20:15

Christmas Conference Dinner
20:30 - 23:00

2021-12-20

M
CFE - CMStatistics
08:15 - 09:30

Coffee Break
09:30 - 10:00

N
CFE - CMStatistics
10:00 - 11:15

O - Keynote
CFE - CMStatistics
11:25 - 12:15

Lunch Break
12:15 - 13:45

P - Keynote
CFE - CMStatistics
13:45 - 14:35

Q
CFE - CMStatistics
14:45 - 16:25

Coffee Break
16:25 - 16:55

R
CFE - CMStatistics
16:55 - 18:35

S - Keynote and closing
CFE - CMStatistics
18:45 - 19:40

VIRTUAL TUTORIALS, MEETINGS AND SOCIAL EVENTS

TUTORIALS

Tutorials will take place on Friday the 17th of December 2021. The first tutorial ("Model selection and inference") will be delivered by Prof. Gerda Claeskens, KU Leuven, Belgium, 9:00-13:30 (GMT). The second tutorial ("Model selection, averaging, shrinkage, and lasso") will be delivered by Prof. Bruce E. Hansen, University of Wisconsin-Madison, US, 15:00 to 19:30 (GMT). Only participants who had subscribed for the tutorial can attend. Registered participants will be able to access the virtual tutorial through the website.

SPECIAL MEETINGS by invitation to group members

- The *Econometrics and Statistics (EcoSta) Editorial Board* and the *CSDA and Annals of Statistical Data Science Editorial Board* meetings will take place on Friday the 17th of December 2021, 15:30-16:00 (GMT).

Indications to attend the virtual Editorial Board meetings will be sent to the AEs attending the conference in due course.

ACCESS TO THE CONFERENCE

- All the participants can attend virtually or in-person, provided that they fulfil the conditions imposed by the UK. However, the in-person access to King's College London for conference participants is restricted to those who had confirmed their in-person participation in the doodle sent by email.
- The in-person venue is King's College London, Strand campus (Strand, London WC2R 2LS, United Kingdom).
- Indication to access the virtual part of the conference can be found on the webpage.

Scientific programme and social events

- The conference is live streaming, and it will not be recorded. The virtual oral presentations will take place through Zoom, while the social events and poster presentations will run in Gather Town.
- **Scientific programme:** The virtual and hybrid sessions are accessible from the interactive schedule. The conference programme time is set in GMT. Indications to access the in-person and virtual rooms can be found on the website. The in-person participants can use S0.13, S2.28, S2.29 and S2.30 as quiet rooms and to participate in virtual sessions with their laptops and headphones.
- **Networking lunch breaks:** During lunchtime each day, the conference participants are invited to interact in the conference virtual networking space. Indications to access the networking space can be found on the website.
- **Welcome reception:** The in-person welcome reception for registered participants will take place at the Chapters/Somerset Rooms of the King's building (Level 2) on Saturday the 18th of December 2021 from 20:00 to 21:30 (GMT). Simultaneously, a virtual welcome reception will take place on Saturday the 18th of December 2021 from 20:00 to 21:30 in Gather Town. Indications to access the can be found on the website.

Presentation instructions

The virtual presentations will take place through Zoom. Speakers should install the application, have a stable internet connection, and ensure their video and audio are working. They will share their slides when the chair requires it, present their talk, and answer the question after the presentation. The in-person speakers must copy their presentations on the desktop on the conference rooms PCs and then share them on Zoom. The PCs have a touch screen with a webcam, a mobile support and an omnidirectional desk microphone that collects the sound around the PC desk to make the live streaming easy. Detailed indications for speakers in either virtual or hybrid sessions can be found on the website. As a general rule, each speaker has 20 minutes for the talk and 3-4 mins for discussion. 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 16th of December. Landscape orientation is advisable. Detailed indications for the poster presentations can be found on the website.

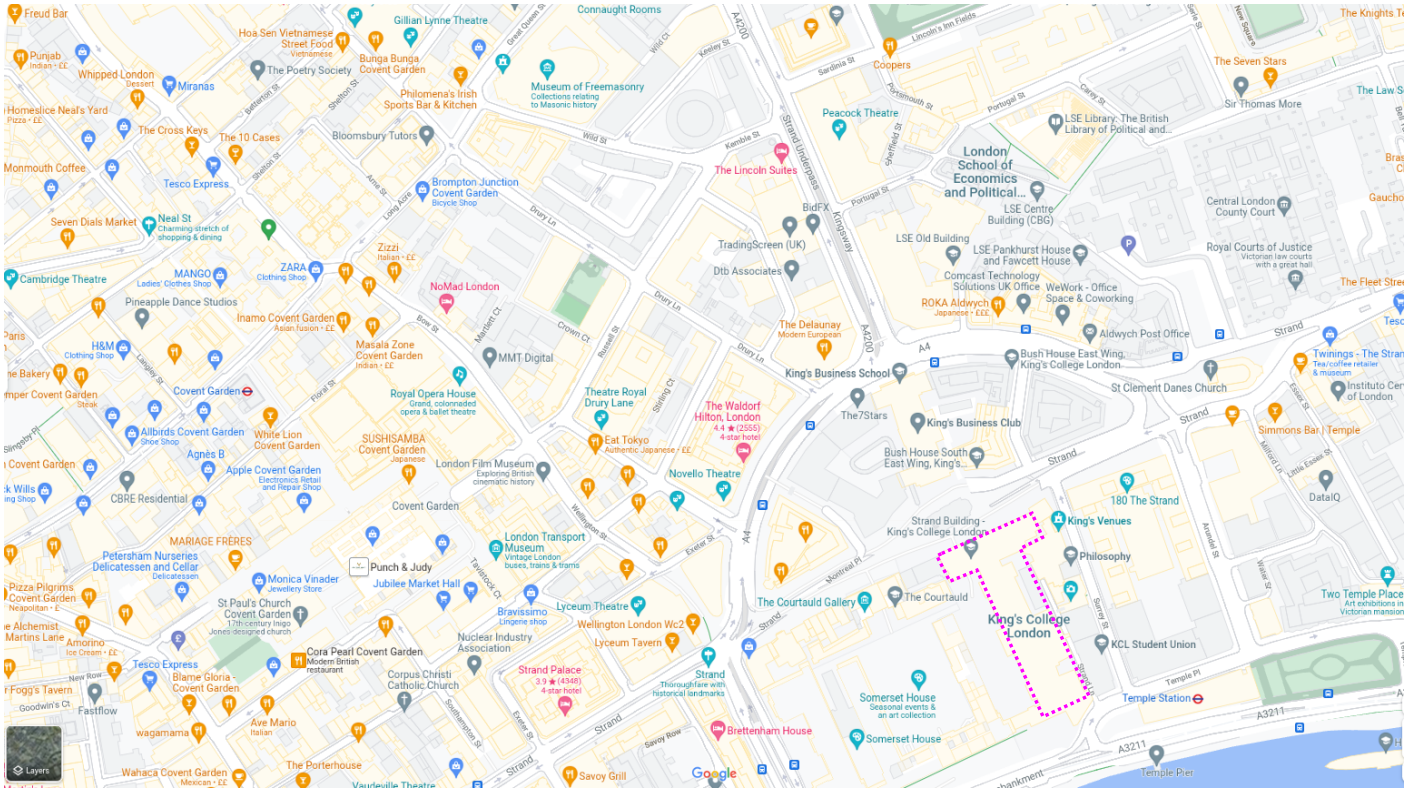
Session chairs

The session chairs will be responsible for introducing the session, the speakers and coordinating the discussion time. A member of the conference staff, identified on Zoom by the name Angel followed by the room number, will assist online. 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 session chairs of both virtual and hybrid sessions can be found on the website.

Test session

A test session will be set up for Saturday the 11th of December 2021 from 14:00 to 15:00 GMT. The participants will be able to enter the virtual Room R18 in the programme to test their presentations, video, micro and audio. Detailed indications for the test sessions can be found on the website.

Map of the venue and nearby area

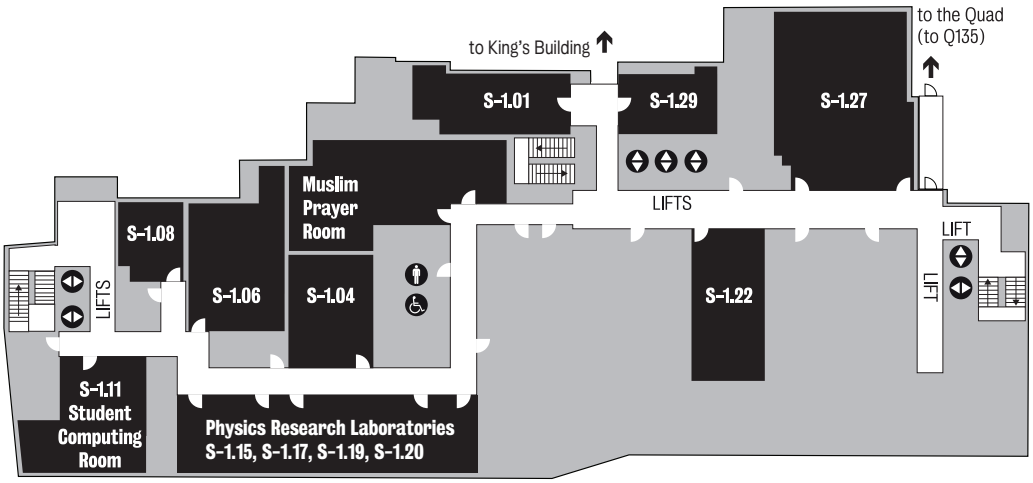


Floor maps

Strand Campus

Strand Building – Basement 1

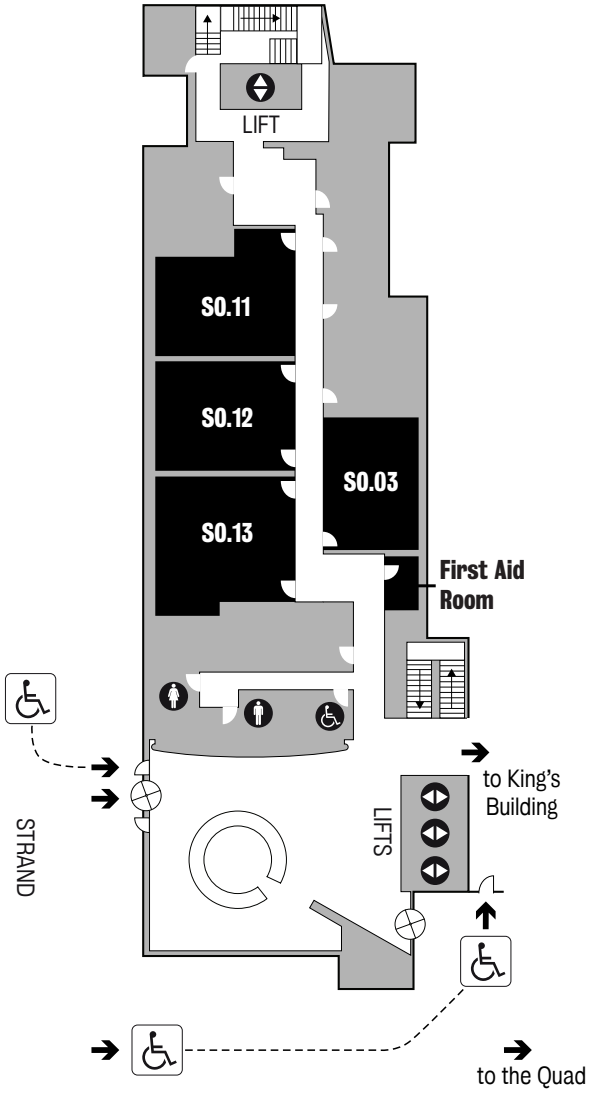
KING'S College LONDON



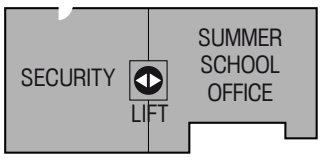
Strand Campus

Strand Building – Ground floor

KING'S College LONDON

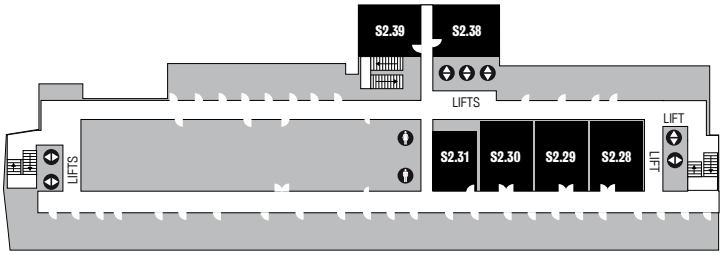


The non-stepped accessible route in to the building is to the rear of the main reception area. This route is via the black gated entrance and turn left.
There is also a button-controlled self-opening door at the front of the main reception but this requires reception staff to activate it.



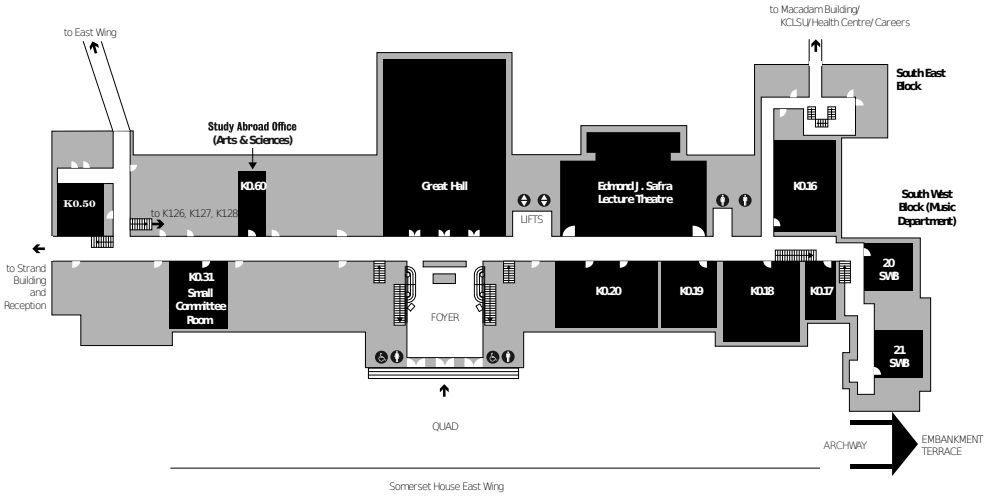
Strand Campus

Strand Building – Floor 2



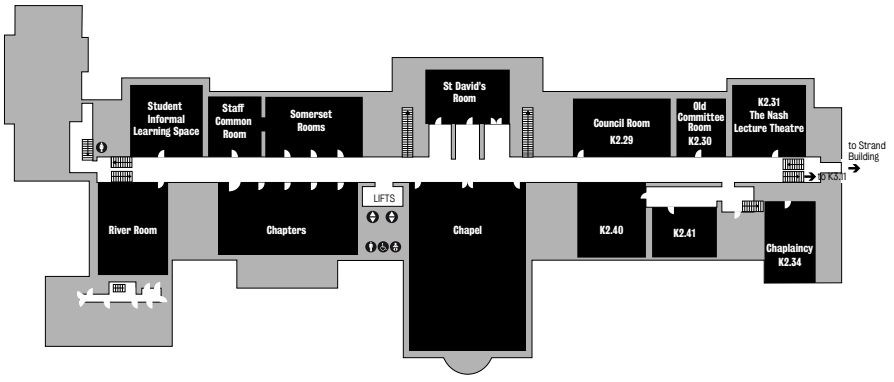
Strand Campus

King's Building – Level 0



Strand Campus

King's Building – Level 2



PUBLICATION OUTLETS

Econometrics and Statistics (EcoSta)

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

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

Call For Papers Econometrics and Statistics (EcoSta)

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

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

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

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

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

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

Editors: Erricos Kontoghiorghes and Ana Colubi (CMStatistics)

Guest Associate Editors: Julyan Arbel, Peter Buhlmann, Stefano Castruccio, Bertrand Clarke, Christophe Croux, Maria Brigida Ferraro, Yulia Gel Michele Guindani, Xuming He, Sangwook Kang, Ivan Kojadinovic, Chenlei Leng, Taps Maiti, Geoffrey McLachlan, Hans-Georg Mueller, Igor Pruenster, Juan Romo, Elvezio Ronchetti, Anne Ruiz-Gazen, Sylvain Sardi, Xinyuan Song, Cheng Yong Tang, Roy Welsch and Peter Winker.

Contents

Saturday 18.12.2021 08:15 - 09:15 Room: K E. Safra (Multi-use 01) Chair: Kalliopi Mylona Keynote talk 1

Monitoring a developing pandemic with available data (virtual)

Speaker: **Jens P. Nielsen, City, University of London, United Kingdom** M. Luz Gamiz, M. Dolores Martinez-Miranda, Enno Mammen

When a pandemic is developing, data collection is chaotic, and the most important data varies from week to week and from country to country. A dynamic approach is provided to monitoring the most critical transitions in a pandemic. The data used are simple, reflecting the type of data almost everybody learned to know during the Covid-19 pandemic. The simplicity of the data is a challenge, and a new missing data methodology has to be developed. The methodology is illustrated via the case of Covid-19 developing in France.

Sunday 19.12.2021 19:25 - 20:15 Room: K E. Safra (Multi-use 01) Chair: Degui Li Keynote talk 2

The exact distribution of the T ratio (virtual)

Speaker: **Bruce Hansen, University of Wisconsin-Madison, United States**

New expressions are presented for the exact finite sample distribution of the heteroskedasticity-robust t-ratio under the assumption of normal heteroskedastic errors. The first expression shows that the distribution function equals the expectation of a nonlinear function of a weighted sum of chi-square random variables, with the weights an explicit function of the regressor matrix and error variances. The second expression shows that the distribution function equals a mixture of student t. distribution functions. These are the first expressions for the exact distribution of the White t-ratio allowing for heteroskedastic error variances, other than expressions based on the numerical inversion of the characteristic function. Our exact distribution function is inconvenient to evaluate in practice, so we recommend a simple approximation with excellent computational and approximation properties. The motivation is the first result described above that the distribution function is completely determined by a specific weighted sum of chi-squares. Using results from the recent literature on the approximation of the distribution function of weighted sums of chi-squares, we obtain a practical approximation to the distribution function of the White t-ratio which is computationally fast in small to moderate samples and is exceedingly accurate.

Monday 20.12.2021 11:25 - 12:15 Room: K E. Safra (Multi-use 01) Chair: Francisco Javier Rubio Keynote talk 3

Multilevel models for longitudinal dyadic family data (in-person)

Speaker: **Fiona Steele, London School of Economics, United Kingdom**

The family is one of the most important examples of a social network. There is substantial interest in studying family interactions to understand child development and, in later life, exchanges of intergenerational support between adult children and their parents. In their simplest form, these data have a dyadic structure with a bivariate response describing bidirectional interactions between the members of each dyad. However, dyadic family data often have a more complex structure: data are increasingly longitudinal, dyads may be nested within families, and a family member may belong to multiple family dyads leading to a cross-classified structure. Multilevel models offer a flexible way of analysing complex longitudinal and multivariate data from dyadic designs. We consider random-effects models for longitudinal dyadic data collected under two different designs: a round robin study of the behaviour observed between each pair of family members over the course of a task, and household panel data on the support that mothers and fathers provide to and receive from their adult children.

Monday 20.12.2021 13:45 - 14:35 Room: K E. Safra (Multi-use 01) Chair: Tommaso Proietti Keynote talk 4

Investigating growth at risk using a multi-country non-parametric quantile factor model (virtual)

Speaker: **Gary Koop, University of Strathclyde, United Kingdom** Todd Clark, Florian Huber, M. Marcellino, Michael Pfarrhofer

A Bayesian non-parametric quantile panel regression model is developed. Within each quantile, the response function is a convex combination of a linear model and a non-linear function, which we approximate using Bayesian Additive Regression Trees (BART). Cross-sectional information at the p^{th} quantile is captured through a conditionally heteroscedastic latent factor. The non-parametric feature of our model enhances flexibility, while the panel feature, by exploiting cross-country information, increases the number of observations in the tails. We develop Bayesian Markov chain Monte Carlo (MCMC) methods for estimation and forecasting with our quantile factor BART model (QF-BART), and apply them to study growth at risk dynamics in a panel of 11 advanced economies

Monday 20.12.2021 18:45 - 19:40 Room: K E. Safra (Multi-use 01) Chair: Cristian Gatu Keynote talk 5

Most powerful inference after model selection via confidence distributions (virtual)

Speaker: **Gerda Claeskens, KU Leuven, Belgium**

Sometimes a model for statistical analysis is not given before the analysis but is the result of a model selection method using the same data. Thus, the uncertainty about the model used for inference has consequences for hypothesis testing and the construction of confidence intervals for the model parameters of interest. Ignoring this uncertainty leads to over-optimistic results, implying that computed p-values are too small and that confidence intervals are too narrow for the intended coverage. Confidence distributions and confidence curves need to be adjusted to account for the selection of the model in order to provide valid inference after model selection for the parameters of interest. Under some assumptions, uniformly most powerful post-selection confidence curves are obtained that are finite sample exact.

Saturday 18.12.2021 09:25 - 10:40 Parallel Session B – CFE-CMStatistics

EO316 Room K E. Safra (Multi-use 01) DESIGN OF EXPERIMENTS (HYBRID) Chair: David Woods

E0325: **General Bayesian design of experiments for calibration of mathematical models**
Presenter: **Antony Overstall**, University of Southampton, United Kingdom

A mathematical model is a representation of a physical system often derived from scientific theory. It is considered to be a function of certain arguments, returning a theoretical prediction of a feature (or features) of the physical system. The arguments are assumed to belong to two groups: (a) controllable inputs, and (b) unknown calibration parameters. Observations of the physical system, for certain controllable inputs, can be used to attribute values to the calibration parameters, a process known as calibration. The values given to the calibration parameters should, in some senses, result in the mathematical model being "close" to the physical system. This goal implicitly recognises that there do not exist values of the calibration parameters such that the mathematical model is equal to the physical system for all values of the controllable inputs. The "distance" between the mathematical model and the physical system can be represented by a loss function, which, in turn, defines a general Bayesian posterior distribution. The aim is to design a calibration experiment, i.e., the choice of controllable inputs at which to observe the physical system to reduce uncertainty exhibited by the general Bayesian posterior under a chosen loss function.

E1162: **Sequential multi-objective planning of factorial experiments**
Presenter: **Olga Egorova**, King's College London, United Kingdom
Co-authors: Steven Gilmour

In numerous experimental settings, especially in the ones pursuing multiple objectives and/or with the lack of preliminary knowledge about the process under study, laying out the whole design at the very beginning is a big decision to make – and in an extensive amount of practical situations, such rigidity is neither necessary nor beneficial. Whether there is a need to run a pilot study first, followed by a series of more focused experimentations, or a situation with the model uncertainty involving combining various, often contradicting optimality criteria. Ideally, we would want to break down the decision-making process, such that each stage is planned (a) making use of the previously obtained data, and (b) accounting for the primary objectives of the following stages. In this most recent work, we explore optimal sequential planning for a series of factorial experiments in the presence of potential model misspecification, so that the quality of the inference from the fitted model is optimised together with minimising the lack-of-fit. We investigate the issues of updating the criteria, models and ensuring the coherence of the results obtained at different stages.

E0480: **Minimax efficient random experimental design strategies with application to model-robust design for prediction**
Presenter: **Tim Waite**, University of Manchester, United Kingdom
Co-authors: David Woods

Fisher stressed the importance of randomizing an experiment via random permutation of the allocation of treatments to experimental units; in an industrial context, this usually amounts to randomizing the run order of the design. We take the idea of experimental randomization much further by introducing flexible new random design strategies in which the design to be applied is chosen at random from a distribution of possible designs. We discuss the philosophical justification for doing so from a game-theoretic perspective and it is shown that the new strategies give stronger bounds on both the expectation and survivor function of the loss distribution. The consequences of this approach are explored in several problems, including global prediction from a linear model contaminated by a discrepancy function from an L_2-class. In this problem the performance improvement is dramatic: the new approach gives bounded expected loss, in contrast to previous designs for which the expected loss was unbounded.

EO523 Room K0.20 (Hybrid 05) PREDICTING AND FORECASTING FOR COMPLEX DATA Chair: Matus Maciak

E1099: **Regularized changepoint detection in panel data models applied for predictions of implied volatility dynamics**
Presenter: **Matus Maciak**, Charles University, Czech Republic

Implied volatility (IV) serves as an important and powerful tool when analyzing financial markets. We propose a novel approach to estimate the overall IV dynamics represented by the underlying panel data model with changepoints. A robust semi-parametric regression framework and atomic pursuit techniques lasso based regularization, in particular, are applied to estimate the underlying analytical structure of the implied volatility surface and a statistical test is used to detect significant changepoints. The overall complexity of the model relies on changepoints that may occur over time, in the analytical structure of the IV smiles, or both. Theoretical and practical details are discussed and the main statistical properties are derived. Empirical properties are investigated in a simulation study and real-life applications are presented to illustration wide and general applicability.

E1210: **Infinitely stochastic micro reserving**
Presenter: **Michal Pesta**, Charles University, Czech Republic
Co-authors: Matus Maciak, Ostap Okhrin

Stochastic forecasting and risk valuation are now front burners in a list of applied and theoretical sciences. We propose an unconventional tool for stochastic prediction of future expenses based on the individual (micro) developments of recorded events. Considering a firm, enterprise, institution, or any entity, which possesses knowledge about particular historical events, there might be a whole series of several related subevents: payments or losses spread over time. This all leads to an infinitely stochastic process at the end. The aim, therefore, lies in predicting future subevent flows coming from already reported, occurred but not reported, and yet not occurred events. The emerging forecasting methodology involves marked time-varying Hawkes process with marks being other time-varying Hawkes processes. The estimated parameters of the model are proved to be consistent and asymptotically normal under simple and easily verifiable assumptions. The empirical properties are investigated through a simulation study. In the practical part of our exploration, we elaborate on a specific actuarial application for micro claims reserving.

E1728: **Reference class selection in similarity-based forecasting of sales growth**
Presenter: **Etienne Theising**, University of Cologne, Germany
Co-authors: Dominik Wied, Daniel Ziggel

A method is proposed to find appropriate outside views for sales forecasts of analysts. The idea is to find reference classes, i.e. peer groups, for each analyzed company separately. Hence, additional companies are considered that share similarities to the firm of interest with respect to a specific predictor. The classes are regarded to be optimal if the forecasted sales distributions match the actual distributions as closely as possible. The forecast quality is measured by applying goodness-of-fit tests on the estimated probability integral transformations and by comparing the predicted quantiles. The method is applied on a data set consisting of 21,808 US firms over the time period 1950 - 2019, which is also descriptively analyzed. It appears that in particular the past operating margins are good predictors for the distribution of future sales. A case study with a comparison of our forecasts with actual analysts estimates emphasizes the relevance of our approach in practice.

EO728 Room Virtual R18 RECENT ADVANCES IN RANDOM MATRIX THEORY AND HIGH DIMENSIONAL STATISTICS Chair: Zeng Li

E0486: **High dimensional linear discriminant analysis: Locally-adaptive shrinkage estimation and false selection rate control**
Presenter: **Bowen Gang**, Fudan University, China
Co-authors: Wenguang Sun

The focus is on the problem of controlling error rate in high dimensional linear discriminant analysis. The problem has two major challenges. First, the oracle Fisher's rule cannot guarantee a low error rate. Second, existing methods for accurate estimation of significance index in high dimensional setting either require strong assumptions that may not hold in practice or have little theoretical support. To address the first challenge, we propose to control a generalization of misclassification rate called false selection rate (FSR). For the second challenge, we propose a locally adaptive shrinkage approach to estimate the class probabilities. In contrast to existing methods, our proposed method does not require the usual sparsity or independence assumptions. The new method is shown to have desirable theoretical properties and reveals an interesting dynamic between estimating discriminant and estimating class probabilities. The numerical performance of the classifier is investigated using both simulated and real data. In particular, the procedure is applied to analyze a lung cancer dataset and is found to perform favorably compared with existing methods.

E1110: **Asymptotics of spatial-sign based estimators of location and scatter in high-dimensions**
Presenter: **Weiming Li**, Shanghai University of Finance and Economics, China

The purpose is to investigate asymptotic behaviors of spatial-sign based location and scatter estimators, i.e., the sample spatial median and its associated spatial-sign covariance matrix, in high-dimensional frameworks. Two stochastic representations for the spatial median are provided with explicit forms, which can characterize the first and second-order fluctuations of the spatial median, in the almost sure sense. Beyond this, a new central limit theorem is established for linear spectral statistics of the spatial-sign covariance matrix. All these results are obtained under a general population model that covers the popular independent components model and the family of elliptical distributions.

E1753: **Distributed PCA for high-dimensional heterogeneous data**
Presenter: **Yanrong Yang**, The Australian National University, Australia

Distributed principal component analysis (DPCA) aims to accurately estimate the principal eigenspace for data stored across multiple local machines. A weighted averaging approach is proposed for DPCA under heterogeneous cases, where the data follow different factor models across local machines but share the same principal eigenspace. Each local machine computes its principal eigenvectors as well as a value "weight", and then transmits them to the central server; the central server aggregates this information from all local machines in a weighted averaging way and conducts PCA based on the aggregated information. Theoretically, we establish the rate of convergence for the weighted averaging DPCA, which demonstrates more efficiency than the previous equal-weight estimator under heterogeneous scenarios (e.g. different sample sizes, factors and error components across local machines). We conduct an extensive simulation study to show the outperformance under various heterogeneities. As a by-product, a new test statistic is proposed to detect the equivalence of principal eigenspace for multiple sets of high dimensional data. We develop the asymptotic distribution for this test statistic and simultaneously apply it in two statistical applications: one is to test change points for daily stock returns while another one is to cluster mortality data from multiple countries.

EO547 Room Virtual R20 STATISTICAL LEARNING IN DECISION MAKING SYSTEMS Chair: Matteo Borrotti

E1357: **Black-box models and Interpretability: Prediction of Italian SMEs' default**
Presenter: **Marco Repetto**, University of Milano-Bicocca, Italy
Co-authors: Lisa Crosato, Caterina Liberati

Assessing Small and Medium Enterprises (SMEs) creditworthiness is a significant issue within organizations. Recent works suggest that SMEs' default prediction is more complex than large enterprises. To handle this complexity, both scholars and practitioners resorted to model credit risk through Machine Learning (ML) techniques, whose applications saw a steep increase in the retail credit risk ambit. However, the lack of interpretability of black-box models has limited their usage in credit risk applications. A possibility of restoring interpretability can be found in reverse-engineering the ML model without accessing its inner parameters, leading to post-hoc explanations that allow the Decision Maker to pin down the relevant effects captured by the black-boxes and decide accordingly. The aim is to model and interpret SMEs' defaults using the eXtreme Gradient Boosting (XGBoost) and the FeedForward Neural Network (FANN) algorithms, and compare them with a few traditional models. We employ recent model-agnostic techniques, such as Accumulated Local Effects and Shapley values, to overcome the usual lack of interpretability of the black-box machines. Results show an overall highest classification power by the XGBoost and highlight the ranking of the input variables based on their contribution to the model outcome as well as the variables impact on the likelihood of default.

E1646: **Statistical modelling for drivers and mediators of the characteristic rest tremor of Parkinson's disease**
Presenter: **Kieran Baker**, King's College London, United Kingdom
Co-authors: Chianna Umamahesan, Clive Weller, Sylvia Dobbs, John Dobbs, Andre Charlett, Steven Gilmour

A global subjective assessment score (MDS-UPDRS) is commonly used clinically to quantify the severity of Parkinson's disease (PD). It is the sum of subscores based on an assessment of a range of features, including tremor. Accelerometers mounted on finger pulp have been used to objectively measure tremor, over the frequency range 3-14 Hz, at rest. Clinicians identified three key metrics, partitioned by frequency band, for tremor analysis: displacement of a device by tremor, total duration of tremor and number of pulses over a 20 second period. Available numerical integration schemes were compared to determine the best for converting acceleration signals to displacement. MDS-UPDRS rest tremor subscores have been evaluated against the corresponding displacement continuous-scale measurement. The discriminatory power of the key tremor metrics for PD-status was tested using logistic regression, with modelling process guided by a clinical understanding of PD. Whilst size of displacement discriminated for PD, the 4-7 Hz rest tremor pattern (characteristic in PD) was also evident in those not diagnosed with PD, potentially quantifying distance down-the-pathway to PD. This aligns with the long pre-presentation state of PD. Statistical and machine learning models will be used to map potential drivers and mediators (faecal metabolome, intestinal inflammation/barrier disfunction, immunome) of tremor.

E0634: **Cluster analysis of diabetic kidney disease longitudinal data**
Presenter: **Veronica Distefano**, Ca Foscari University of Venice, Italy
Co-authors: Maria Mannone, Claudio Silvestri, Irene Poli

Information on patients' heterogeneity of disease progress is crucial while seeking for individualized therapeutic treatments in the framework of precision medicine. We focus on a diabetic kidney disease dataset, clustering patients according to their eGFR (estimated glomerular filtration rate) trajectories. We find subgroups of similar patients with similar time trajectories. We also compare given drug combinations with individual responses to the treatments. To draw information on shape similarity of time trajectories, we apply the Frechet distance, grouping similar curves in clusters, highlighting time characteristics for each cluster.

EO832 Room Virtual R21 JOINT MODELLING FOR LONGITUDINAL AND SURVIVAL DATA Chair: Ipek Guler

E0944: **The joint models for longitudinal and survival data: Analysis and extensions**
Presenter: **Marcella Mazzoleni**, University of Bergamo, Italy

The joint models for survival and longitudinal data became an appealing topic these years. In fact, several researchers decided to focus and to propose possible extensions of these models. The original idea of joint models was to jointly analyse the two sub-models, namely longitudinal and survival, quantifying the effect of one longitudinal covariate on the risk of an event. Starting from this point different extensions of the sub-models, several estimation methods, and various applications, most of which in the medical field, were proposed. The focus is on extending the longitudinal sub-model, analysing more than just one longitudinal covariate, appropriately adjusting the estimation method based on maximising the likelihood function through the implementation of an Expectation-Maximisation algorithm. Accordingly, the diagnostic and goodness of fit

elements are updated considering more than just one longitudinal covariate, implementing the estimated survival function, and the residuals and dynamic predictions for both survival and longitudinal sub-models.

E1428: **Extensions on joint models for multivariate non-linear longitudinal and survival data**
Presenter: **Ipek Guler**, KU Leuven, Belgium

Many follow up studies in biomedical research produce both repeated measurements and time-to-event analysis and, commonly, it is of interest to explore the association between them. For this aim, many statistical developments have been proposed to jointly model longitudinal and survival data. Further from an association between a single longitudinal biomarker and survival data, there are many extensions on multivariate longitudinal and multivariate survival data using either maximum likelihood or Bayesian approaches. However, in many situations, the computational cost is getting higher in case of having an increased number of outcomes so then the dimensional increase on the random effects covariance matrix. We will focus on the pairwise approach for multivariate longitudinal data and its use on joint modelling extensions. We will illustrate the different model approaches proposed in the literature with real biomedical data.

E1431: **Novel joint modelling extensions for survival and longitudinal data applied to COVID-19**
Presenter: **Carla Diaz Louzao**, Universidade de Santiago de Compostela, Spain
Co-authors: Ipek Guler, Francisco Gude, Santiago Tome, Carmen Cadarso Suarez

Since the notification at the end of 2019 of several cases of pneumonia caused by the coronavirus SARS-CoV-2, the expansion of the infection has been very fast worldwide, and nowadays it is considered pandemic. The main affected organ is the lung, but liver failure is frequent during the infection, with an elevation in transaminase levels related with the severity of the disease. Furthermore, numerous studies have reported that higher levels of inflammation markers are associated with higher rates of mortality. It is also suspected that these markers influence significantly in the transaminase levels. With this in mind, the longitudinal inflammation and transaminase levels and the risk of death for patients hospitalized in the University Clinical Hospital (Santiago de Compostela, Spain), during the first wave of the coronavirus disease in Spain, are jointly modelled. The joint modelling approaches for longitudinal and survival data have been widely used in follow-up studies in to explore the relationship between longitudinal biomarkers and survival. The novel extensions on joint models are based on the multivariate longitudinal and survival data. One of these extended models is applied to explore the association between the inflammation markers, transaminase levels and risk of death on SARS-CoV-2.

EO583 Room Virtual R22 LIMIT THEOREMS FOR STOCHASTIC PROCESSES Chair: Salim Bouzebda

E1075: **Renewal type bootstrap for U- process Markov chains**
Presenter: **Inass Soukarieh**, Universita de Technologie de Compiegne, France

The main purpose is to establish bootstrap uniform functional central limit theorems U-processes for Harris recurrent Markov chains over uniform classes of functions satisfying some entropy condition. To simplify our approach, we will make use of the well-known regenerative properties of Markov chains avoiding some complicated mixing conditions. We show that the U-processes satisfies the bootstrap uniform CLT with minimal condition on envelope function. We next consider an extension to the k Markov chains setting and prove the bootstrap consistency. The theoretical uniform central limit theorems set out are (or will be) key tools for many further developments in Markovian data analysis.

E1102: **Uniform in number of neighbors consistency for the conditional U-statistics involving functional data**
Presenter: **Amel Nezzal**, Universite de Technologie de Compiegne, France
Co-authors: Salim Bouzebda

U-statistics represent a fundamental class of statistics arising from modeling quantities of interest defined by multi-subject responses.U-statistics generalise the empirical mean of a random variable X to sums over every m-tuple of distinct observations of X. The class of so-called conditional U-statistics may be viewed as a generalization of the Nadaraya-Watson estimates of a regression function. We introduce the k nearest neighborhoods estimator of the conditional U-statistics and establish uniform in $\mathbf{t}$ and in the number of neighborhoods (UINN) (at some specific rate) to $m(\mathbf{t})$ when Y and covariates X are functional taking value in some abstract spaces. In addition, uniform consistency is also established over $\varphi \in \mathcal{F}$ for a suitably restricted class $\mathcal{F}$ in both cases bounded and unbounded satisfying some moment conditions. The approaches in some recent papers are unified. The theoretical uniform consistency results are (or will be) key tools for many further developments in functional data analysis. The theorems allow data-driven local bandwidths for conditional estimators

E1119: **General M-estimator processes and their m out of n bootstrap with functional nuisance parameters**
Presenter: **Anouar Abdeldjaoued Ferfache**, University de Technologie de Compiegne, France
Co-authors: Salim Bouzebda

The focus is on the problem of the estimation of a parameter theta, in Banach spaces, maximizing some criterion function which depends on an unknown nuisance parameter h, possibly infinite-dimensional. Classical estimation methods are mainly based on maximizing the corresponding empirical criterion by substituting the nuisance parameter with some nonparametric estimator. We show that the M-estimators converge weakly to maximizers of Gaussian processes under rather general conditions. The conventional bootstrap method fails, in general, to consistently estimate the limit law. We show that the m out of n bootstrap, in this extended setting, is weakly consistent under conditions similar to those required for weak convergence of the M-estimators. The aim is therefore to extend the existing theory on the bootstrap of the M-estimators. Examples of applications from the literature are given to illustrate the generality and usefulness of the results. Finally, we investigate the performance of the methodology for small samples through a short simulation study.

EO738 Room Virtual R23 RECENT ADVANCES IN LATENT VARIABLE MODELS Chair: Roberto Di Mari

E0310: **Latent Markov factor analysis for evaluating measurement model heterogeneity in intensive longitudinal data**
Presenter: **Leonie Vogelsmeier**, Tilburg University, Netherlands
Co-authors: Jeroen Vermunt, Kim De Roover

When studying intensive longitudinal data (e.g., with Experience Sampling Methodology), drawing conclusions about dynamics of psychological constructs (e.g., well-being) over time requires the measurement model (MM; indicating which items measure which constructs) to be invariant between subjects and within subjects over time. However, there might be heterogeneity or non-invariance in the MM, for instance, due to subject-specific differences and changes in item interpretation or response styles. Mixture modeling approaches have proved to be powerful tools to detect unobserved heterogeneity, but the methodology to evaluate measurement invariance for multiple time points and subjects simultaneously was lacking. To fill this gap, we built upon common mixture modeling approaches and proposed latent Markov factor analysis (LMFA), which combines a discrete- or continuous-time latent Markov model (that clusters observations into separate states, according to state-specific MMs) with mixture factor analysis (that evaluates which MM applies for each state). We introduce this novel methodology, illustrate it by means of an empirical example, discuss two possible estimation procedures, and explain the latest extension, latent Markov latent trait analysis (LMLTA), that adequately deals with ordinal responses.

E0673: **Selecting clustering algorithms and solutions via quadratic scoring**
Presenter: **Luca Coraggio**, University of Naples Federico II, Italy
Co-authors: Pietro Coretto

A novel methodology is introduced to score clustering solutions and select the optimal one from a set of candidate solutions. In particular, we develop a framework where clustering solutions are represented via triplets of parameters of clusters' proportions, centres and scatters; this representation is used together with the quadratic score function (central to Quadratic Discriminant Analysis) to develop two novel cluster quality criteria, named quadratic scores. These assess the extent to which sample points are well accommodated into quadratic regions defined by the clustering. We show that the proposed criteria are consistent with clusters generated from a restricted class of mixtures of elliptical-symmetric distributions, including the Gaussian model. Nonetheless, the proposed criteria are method-independent: they do not rely on any particular clustering framework or algorithm and can be computed for any clustering solution. We also propose variations on the quadratic scores, which make use of cross-validation and bootstrap resampling. We compare our proposals with several established criteria from the literature, used to select clustering solutions; these include method-independent and model-based criteria. The proposed methodology proves to achieve among the highest performances in an extensive empirical study on both simulated and real data sets, involving 440 clustering solutions per data set.

E1021: **State-switching varying-coefficient stochastic differential equations**
Presenter: **Timo Adam**, University of St Andrews, United Kingdom
Co-authors: Richard Glennie, Theo Michelot

Stochastic differential equations (SDEs) are popular tools for uncovering mechanistic relationships underlying time series data. By modelling the parameters of the process of interest as potentially smooth functions of a given set of covariates, varying-coefficient SDEs provide an extension of basic SDEs that allows us to capture more detailed, non-stationary features of the data-generating process. However, in practice, these parameters often vary at multiple time scales, which will be illustrated using dive data of Bairds beaked whales: while changes in pitch, roll, and heading exhibited within some dives can be described by some varying-coefficient SDE, other dives can be better characterised by other varying-coefficient SDEs; a pattern that is not readily accommodated for by the existing approach. To account for such state-switching patterns between dives while simultaneously allowing to make inference on the underlying behavioural processes that occur within dives, we propose a Markov chain operating at the between-dives scale that selects among a finite set of varying-coefficient SDEs to model the behaviour at the within-dive scale. The resulting class of state-switching varying-coefficient SDEs thus allows us to simultaneously model time-series data at multiple time scales in a joint modelling framework.

EO796 Room Virtual R24 MACHINE LEARNING AND STATISTICAL INVERSE PROBLEMS Chair: Catia Scricciolo

E0702: **Equivariant neural networks for inverse problems**
Presenter: **Ferdia Sherry**, University of Cambridge, United Kingdom

In recent years, the use of convolutional layers to encode an inductive bias (translational equivariance) in neural networks has proven to be a very fruitful idea. The successes of this approach have motivated a line of research into incorporating other symmetries into deep learning methods, in the form of group equivariant convolutional neural networks. Much of this work has been focused on the roto-translational symmetry of Euclidean spaces, but other examples are the scaling symmetry of Euclidean spaces and the rotational symmetry of the sphere. We demonstrate that group equivariant convolutional operations can naturally be incorporated into learned reconstruction methods for inverse problems that are motivated by the variational regularisation approach. Indeed, if the regularisation functional is invariant under a group symmetry, the corresponding proximal operator will satisfy an equivariance property with respect to the same group symmetry. As a result of this observation, we design learned iterative methods in which proximal operators are modelled as group equivariant neural networks. We use roto-translationally equivariant operations in the proposed methodology and apply it to the problems of low-dose CT reconstruction and subsampled MRI reconstruction. The proposed methodology is demonstrated to improve the reconstruction quality of a learned reconstruction method with a little extra computational cost at training time but without any extra cost at test time.

E1284: **Efficient semiparametric estimation and cut posterior contraction in semiparametric hidden markov models**
Presenter: **Daniel Moss**, University of Oxford, United Kingdom
Co-authors: Judith Rousseau

The problem of estimation in hidden markov models with finite state space and nonparametric emission distributions is considered. Efficient estimators for the transition matrix are exhibited, and a semiparametric bernstein-von mises result is deduced, extending existing work for mixture models. Following from this, a cut posterior approach is employed to jointly estimate the transition matrix and the emission distributions. A general theorem on contraction rates for such cut posterior approaches is derived, analogous to existing results for the bayesian posterior. This result is applied to obtain a contraction rate result for the emission distributions in our setting, by first proving an L^1 inversion inequality to go from marginals to emissions. Finally, simulation studies are provided to illustrate theoretical results.

E1655: **Bayesian Wasserstein deconvolution**
Presenter: **Catia Scricciolo**, University of Verona, Italy
Co-authors: Judith Rousseau

The focus is on the problem of recovering a distribution function from independent replicates additively contaminated with random errors whose distribution is known to the observer who can record only noisy observations. We investigate whether a Bayesian nonparametric approach for modelling the latent distribution may yield inferences with frequentist asymptotic validity under the 1-Wasserstein metric. When the error density is ordinary smooth, we develop an inversion inequality relating the L^1-distance between mixtures to the 1-Wasserstein distance between the corresponding mixing distributions. This inequality improves on the existing ones when no assumption on the mixing distribution, except for moment constraints, is postulated and yields information-theoretic optimal rates. In fact, minimax-optimal posterior contraction rates for the mixed densities yield optimal rates for the corresponding mixing distributions. An application of this inversion inequality to the deconvolution problem shows that, when the mixing distribution is Lebesgue absolutely continuous, a careful choice of the prior law acting as an efficient approximation scheme for the sampling density leads to a posterior contraction rate equal, up to a log-factor, to the lower bound estimation rate. The same prior law is shown to also adapt to the regularity level of a mixing density belonging to a Sobolev space, thus leading to a new adaptive estimation method with respect to the Wasserstein loss.

EO557 Room Virtual R25 THE ROLE OF BIOSTATISTICS FOR EPIDEMIOLOGIC DESIGNS AND ANALYSES Chair: Paola Rebora

E0515: **Treating ordinal outcomes as continuous quantities: When, why and how**
Presenter: **Chuen Seng Tan**, Saw Swee Hock School of Public Health, Singapore
Co-authors: Yilin Ning, Peh Joo Ho, Nathalie Stoer, Ka Keat Lim, Hwee-Lin Wee, Mikael Hartman, Marie Reilly

Ordinal variables are common in studies of patient care. Analysing such outcomes often utilizes the linear regression model to estimate the effect of an exposure or intervention of interest. The magnitude of the effect is quantified by the difference in mean ordinal scores of the two groups being compared, and this quantity is useful for the assessment of clinical significance. However, this approach may be inappropriate as it assumes the ordinal outcome is a proxy for the continuous scale but does not assess this assumption. We propose a new procedure using the cumulative link model to assess the proxy assumption and to estimate the difference in mean ordinal scores when appropriate. The procedure is applied to 5 subscales of fatigue measured using the Multidimensional Fatigue Inventory to investigate the effect of time since diagnosis on fatigue among breast cancer survivors. A statistically significant improvement over time since cancer diagnosis was found in the General Fatigue and Mental Fatigue

scores, but only General Fatigue satisfied the proxy assumption. We can only draw conclusions on the magnitude of change in the General Fatigue score, which is expected to be 1-unit for every 6.5 additional years since diagnosis and clinical significance (i.e., a 2-unit difference) achieved at the 13-th year. The procedure offers a seamless way to assess both the statistical and clinical significance of an effect on ordinal outcomes when the proxy assumption is appropriate.

E0853: **Weighted analyses of survival outcomes under complex study designs: An R implementation**
Presenter: **Yilin Ning**, National University of Singapore, Singapore

The nested case-control (NCC) design is widely used in survival studies when full cohort analysis is not feasible. NCC samples are often analyzed using conditional logistic regression models, but weighted analyses of sampled subjects after breaking the matching bring additional benefits, e.g., improved statistical efficiency, and estimation of absolute risk and effects of matching factors (if any). Incorporation of sampling weights in analyses becomes essential to the unbiased and efficient estimation of exposure effects under some more complex study designs, e.g., counter-matching for studying rare exposures and extreme case-control design that maximizes information given limited sample size by focusing on early deaths and longest survivors. Benefits of these designs and weighted analyses have been demonstrated in clinical examples, but the uptake of these approaches remains low in practice, partly due to the lack of software tools to conveniently compute the required weights. The SamplingDesignTool R package (https://github.com/nyilin/SamplingDesignTools) closes the gap between methodological developments and practical applications by providing simple commands for drawing samples and computing sampling weights under aforementioned designs, supporting external approximations to risk set sizes when the full cohort is unavailable. The use of the package and the benefit of weighted analyses under the various designs will be illustrated using simulated data.

E1190: **Design2phase library to estimate power and efficiency of a two-phase design with survival outcome**
Presenter: **Francesca Graziano**, University of Milano-Bicocca, Italy
Co-authors: Paola Rebora

The availability of large epidemiological cohorts and stored biological specimens allows re-use these data to answer new research questions. Two-phase sampling is a general approach for sub-sampling that significantly reduce the time and cost of the study. The lack of easily available tools in the selection of the most efficient and powerful sub-cohort is, however, one of the main limitations. The design2phase library, implemented in R software, is a tool that provides a simulation-based investigation of sub-sampling performances with the aim of estimating the association between a new marker and a time-to-event outcome in a two-phase study. The library was created to estimate the power and efficiency of a wide variety of sampling designs (simple random sampling, case-control, probability proportional to size, nested case-control, and countermarching) applying a two-phase Cox model weighted by the inverse of the empirical inclusion probability. This user-friendly tool also gives the possibility to perform stratified sampling and to visualize power curves, and therefore it could be used by the researchers during the planning phase. The statistical background is briefly reviewed and the functions are illustrated on real data on childhood acute lymphoblastic leukemia, to evaluate the role of different genetic polymorphisms on treatment failure due to relapse.

EO826 Room Virtual R26 EXTREMES AND CAUSALITY Chair: Valerie Chavez-Demoulin

E0251: **Detection of causality in time series using extreme values**
Presenter: **Juraj Bodik**, UNIL Lausanne, Switzerland

We deal with the following problem: Let us have two stationary (possibly nonlinear) time series with heavy-tailed marginal distributions. We want to detect whether there is some Granger causality present. Even more, we want to determine the minimal lag, i.e. the time how much it takes for information to travel from one time series to another. We will examine the asymmetry in extremes between the cause and effect, and present a statistic that can estimate such asymmetries. The basis of the idea stands by the so-called causal tail coefficient for time series, which in some way represents the behaviour in extremes of one series conditioned on the presence of an extreme in the other.

E0872: **Estimating an extreme Bayesian network via scalings**
Presenter: **Mario Krali**, EPFL, Switzerland
Co-authors: Claudia Klueppelberg

A recursive max-linear vector models causal dependence between its components by expressing each node variable as a max-linear function of its parental nodes in a directed acyclic graph and some exogenous innovation. Motivated by extreme value theory, innovations are assumed to have regularly varying distribution tails. We propose a scaling technique in order to determine a causal order of the node variables. All dependence parameters are then estimated from the estimated scalings. Furthermore, we prove asymptotic normality of the estimated scalings and dependence parameters based on the asymptotic normality of the empirical spectral measure. Finally, we apply our structure learning and estimation algorithm to financial data and food dietary interview data.

E1689: **Extremal quantile treatment effects for heavy-tailed distributions**
Presenter: **Sebastian Engelke**, University of Geneva, Switzerland
Co-authors: David Deuber, Marloes Maathuis, Jinzhou Li

Causal inference for rare events has important applications in many fields such as medicine, climate science and finance. We introduce an extremal quantile treatment effect as the difference of extreme quantiles of the potential outcome distributions. Estimation of this effect is based on extrapolation results from extreme value theory in combination with a new counterfactual Hill estimator that uses propensity scores as adjustment. We establish the asymptotic theory of this estimator and propose a variance estimation procedure that allows for valid statistical inference. Our method is applied to analyze the effect of college education on high wages.

EO104 Room Virtual R28 NEW METHODS AND MODELS FOR ORDINAL AND MIXED-TYPE DATA Chair: Cristina Mollica

E1006: **Finite mixtures of discretized beta distribution to model polarization and floatation of ordinal data**
Presenter: **Rosaria Simone**, University of Naples Federico II, Italy

Flexible mixture modelling based on the discretized Beta distribution is proposed to parameterize polarization and floatation of ordered discrete evaluations, as ratings and count data. The framework entails versatile interpretation of results and broad applicative outreach in all studies where the interest lies in disclosing and characterizing clusters of opposite and intermediate responses. A summarizing discussion with respect to identifiability issues is presented. Finally, the proposal is supported with empirical evidence from several datasets, assuming a comparative perspective with alternative approaches.

E1077: **Modeling ranking data in a social network**
Presenter: **Philip Yu**, The Education University of Hong Kong, Hong Kong
Co-authors: Jiaqi Gu

Human interaction and communication have become one of the essential features of social life. Individuals preference behaviors may be influenced from those of their peers or friends in a social network. However, most existing statistical models and methods for ranking data assume independence among the rank-order preferences of different individuals. We introduce a new class of probabilistic models for ranking data in a social network. The new models are able to account for social dependencies among the individuals. An efficient MCMC algorithm is developed for Bayesian inference. Simulation and empirical studies reveal the usefulness of our proposed methods.

E1595: **Efficient estimation of finite mixtures of Mallows models with the Spearman distance**
Presenter: **Cristina Mollica**, Sapienza Universita di Roma, Italy
Co-authors: Marta Crispino, Valerio Astuti, Luca Tardella

The class of Mallows models (MMs) occupy a central role in the literature for the analysis and learning of preferences from a sample of ranking data. The MMs rely on the distance notion over the set of permutations but, despite the wide range of possible metrics, the choice is typically limited to the Kendall or Cayley distances, due to the related analytical simplifications. We go beyond these conventional few options and explore the formal properties of the MM with the Spearman distance, also referred to as the theta-model. The attractive feature of this model is its correspondence with the restriction of the normal distribution over the permutation set such that, similarly to the Gaussian density, the theta model enjoys a convenient closed-form expression for the critical estimation of the modal ranking. This means that, differently from the MMs with the other metrics, an efficient and accurate inferential procedure can be developed, where the computational burden of inferring the discrete parameter is significantly reduced. Additionally, an efficient estimation within the finite mixture framework is realized via the EM algorithm, for enlarging the applicability of theta-models to samples of rankings characterized by a group structure. Finally, an application to a real-world dataset endorsing our proposals in the comparison with competing mixtures of ranking models is provided.

EO842 Room Virtual R36 ADVANCES IN DEPTH AND QUANTILE METHODS Chair: Germain Van Bever

E0340: **Halfspace depth revisited**
Presenter: **Petra Laketa**, Charles University, Czech Republic
Co-authors: Stanislav Nagy, Dusan Pokorny

A halfspace depth of a given point with respect to a probability measure is defined as the infimum of the probabilities of all the closed halfspaces that contain that point. As such, halfspace depth measures the centrality of points with respect to a given probability measure and is therefore used as a multivariate quantile. The existing literature on this interesting topic usually imposes restrictive assumptions on measure. We consider halfspace depth in a general setting, for all finite Borel measures, with the intention to collect partial results from the literature and give more general theoretical results. We specially focus on 1) when and how is it possible to reconstruct the underlying measure based on its halfspace depth function and 2) extending the so-called ray basis theorem, which gives an interesting characterization of the point with the maximal halfspace depth, called the halfspace median.

E0524: **Approximate computation of projection depths**
Presenter: **Pavlo Mozharovskyi**, Telecom Paris, Institut Polytechnique de Paris, France
Co-authors: Rainer Dyckerhoff, Stanislav Nagy

Data depth is a concept in multivariate statistics that measures the centrality of a point in a given data cloud in a Euclidean space. If the depth of a point can be represented as the minimum of the depths with respect to all one-dimensional projections of the data, then the depth satisfies the so-called projection property. Such depths form an important class that includes many of the depths that have been proposed in the literature. For depths that satisfy the projection property, an approximate algorithm can easily be constructed since taking the minimum of the depths with respect to only a finite number of one-dimensional projections yields an upper bound for the depth with respect to the multivariate data. Such an algorithm is particularly useful if no exact algorithm exists or if the exact algorithm has a high computational complexity, as is the case with the halfspace depth or the projection depth. To compute these depths in high dimensions, the use of an approximate algorithm with better complexity is surely preferable. Instead of focusing on a single method, we provide a comprehensive and fair comparison of several methods, both already described in the literature and original.

E1526: **Spatial quantiles on the hypersphere**
Presenter: **Dimitri Konen**, Universita Libre de Bruxelles, Belgium
Co-authors: Davy Paindaveine

A concept of quantiles for distributions on the unit hypersphere R^d is proposed. The innermost quantiles are Frechet medians, i.e. the L^1-analog of Frechet means. Since these medians may be non-unique, we define a quantile field around each such median m. The corresponding quantiles are directional in nature: they are indexed by a scalar order between 0 and 1 and a unit vector in the tangent space to the hypersphere at the median. To ensure computability in any dimension, our quantiles are essentially obtained by considering the Euclidean Chaudhuri spatial quantiles in a suitable stereographic projection of the hypersphere onto its tangent space at the median. Despite this link with their Euclidean antecedent, studying our quantiles requires understanding the nature of the Chaudhuri quantile in a version of the projective space where all points at infinity are identified. We thoroughly investigate the properties of the proposed quantiles, and study in particular the asymptotic behaviour of their sample versions, which requires controlling the impact of estimating the median. Our spherical quantile concept also allows for companion concepts of ranks and depth on the hypersphere.

EO814 Room Virtual R37 MULTIVARIATE TIME SERIES MODELING Chair: Efstathia Bura

E0599: **A test for the number of factors in dynamic factor models**
Presenter: **Daniel Pena**, Universidad Carlos III de Madrid, Spain

An eigenvalue ratio test for the number of dynamic factors is presented. The test combines the advantages of those proposed previously for the eigenvalues of the covariance matrix, and for the cumulative sum of lagged covariance matrices. A pooled correlation matrix is defined as a weighted combination of the main observed correlation matrices and the proposed test is based on the ratio of consecutive eigenvalues. Some theoretical results are given to justify the good expected properties of the test and a Monte Carlo study is presented showing its good finite sample performance compared to previous approaches. The usefulness of the test is also illustrated in an example with real macroeconomic data.

E0644: **A state-space approach to time-varying reduced rank regression**
Presenter: **Barbara Brune**, TU Wien, Austria
Co-authors: Wolfgang Scherrer, Efstathia Bura

A new approach is proposed to reduced-rank regression that allows for time-variation in the regression coefficients. The Kalman filter-based estimation allows for the usage of standard methods and easy implementation of our procedure. The EM algorithm ensures convergence to a local maximum of the likelihood. Our estimation approach in time-varying reduced-rank regression performs well in simulations, with an amplified competitive advantage in time series that experience large structural changes. We illustrate the performance of our approach with a simulation study and two applications to stock index and Covid-19 case data.

E1242: **Factor analysis for data with heterogenous blocks**
Presenter: **Tatyana Krivobokova**, University of Vienna, Austria

It has been often empirically observed that including more variables into factor-based forecasting models may worsen the prediction considerably. We examine this issue assuming a factor model, which consists of heterogeneous and possibly dependent blocks of variables. We identify settings that cause the poor forecasting performance of such factor models estimated by principle component analysis (PCA) and suggest a simple modification of the standard PCA, called blocked PCA (bPCA), that leads to the proper identification of factors and prediction. A simulation study and real data analysis illustrate our findings.

EO493 Room Virtual R38 METHODS FOR FUNCTIONAL TIME SERIES Chair: Juhyun Park

E0418: **On the estimation of nonstationary functional data**
Presenter: **Daisuke Kurisu**, Tokyo Institute of Technology, Japan
An asymptotic theory is developed for estimating the time-varying characteristics of locally stationary functional time series. We introduce a kernel-based method to estimate the time-varying covariance operator and the time-varying mean function of a locally stationary functional time series. Subsequently, we derive the convergence rate of the kernel estimator of the covariance operator and associated eigenvalue and eigenfunctions. We also establish a central limit theorem for the kernel-based locally weighted sample means. As applications of our results, we discuss the prediction of locally stationary functional time series and methods for testing the equality of time-varying mean functions in two functional samples.

E0626: **On lagged covariance and cross-covariance operators of processes in cartesian products of abstract Hilbert Spaces**
Presenter: **Sebastian Kuehnert**, WINGAS GmbH, Germany
A key concern in Functional Time Series Analysis is measuring the dependence within and between processes, for which lagged covariance and cross-covariance operators have proven to be a practical tool. Probabilistic features of and estimators for lagged covariance operators of stationary processes with values in $L^2[0,1]$, the space of measurable, square-Lebesgue integrable real-valued functions with domain [0,1], are widely studied for fixed lag, and under several limitations also in further spaces. Lagged cross-covariance operators of stationary processes in $L^2[0,1]$ were also comprehensively studied. Core results on lagged covariance and cross-covariance operators of processes in cartesian products of abstract Hilbert spaces are reviewed. Motivating examples for the use of processes in such Cartesian products are given. Estimators and asymptotic upper bounds of the estimation errors for the lagged covariance and cross-covariance operators of processes in these Cartesian products are deduced for fixed as well as increasing lag and Cartesian powers. The processes are allowed to be non-centered, and to have values in different spaces when investigating the dependence between processes. Also, estimators for the principle components of our covariance operators are discussed, and a simulation study is performed on well-established time series.

E1094: **Simultaneous predictive bands for functional time series using minimum entropy sets**
Presenter: **Jairo Cugliari**, Universita Lumiere Lyon 2, France
Co-authors: Nicolas Hernandez, Julien Jacques
Functional Time Series (FTS) are sequences of dependent random elements taking values on some functional space. Most of the research on this domain is focused on producing a predictor able to forecast the value of the next function having observed a part of the sequence. For this, the Autoregressive Hilbertian process is a suitable framework. We address the problem of constructing simultaneous predictive confidence bands for a stationary FTS. The method is based on an entropy measure for stochastic processes, in particular FTS. To construct predictive bands, we use a functional bootstrap procedure that allows us to estimate the prediction law through the use of pseudo-predictions. Each pseudo-realisation is then projected into a finite-dimensional space associated with a functional basis. We use Reproducing Kernel Hilbert Spaces (RKHS) to represent the functions, considering then the basis associated with the reproducing kernel. Using a simple decision rule, we classify the points on the projected space among those belonging to the minimum entropy set and those that do not. We push back the minimum entropy set to the functional space and construct a band using the regularity property of the RKHS. The proposed methodology is illustrated through artificial and real-world data sets.

EC877 Room Virtual R29 CONTRIBUTIONS IN COPULAS Chair: Roel Braekers

E1598: **Inference for copulas with two-piece margins**
Presenter: **Anneleen Verhasselt**, Hasselt University, Belgium
Copulas provide a versatile tool in the modelling of multivariate distributions. With increased awareness of possible asymmetry in data, skewed copulas combined with classical margins have been employed to model these data appropriately. The reverse, skewed margins with a (classical) copula has also been considered, but mainly with skew-symmetrical margins. We focus on different types of skewed margins, namely the two-piece distributions. More specifically, we use the recently proposed quantile-based asymmetric family of distributions in given copula structures. For this combination, we provide statistical inference results in consistency and asymptotic normality for the Inference Functions for Margins estimator. A simulation study complements the theoretical results, and the practical usefulness is shown through some real data examples.

E1676: **Geometric-extreme stable distributions and Archimedean copulas**
Presenter: **Violetta Piperigou**, University of Patras, Greece
A family of copulas is introduced by considering the joint distribution of the maximum (and/or the minimum) of two random samples of the same random number of continuous random variables. This family has as a parameter the Laplace transform of a non-negative random variable and in addition a real-valued parameter. It can be seen that the marginal distributions of this joint distribution is an extension of the geometric-extreme stable distributions and also that the family of Archimedean Copulas, with the generator of the inverse of a Laplace transform, can be obtained as a limiting case of the real-valued parameter of the new family of copulas. Various properties of this family are discussed and a special case is studied in detail.

E1107: **Bivariate vine based quantile regression**
Presenter: **Marija Tepegjozova**, Technical University Munich, Germany
Co-authors: Claudia Czado
The statistical analysis of univariate quantiles is a well developed and researched topic. However, there is a profound need for research in multivariate quantiles. We tackle the topic of bivariate quantiles and bivariate quantile regression using vine copulas. They are graph theoretical models composed of a sequence of linked trees, which allow for separate modeling of marginal distributions and the dependence structure in the data. We introduce a novel graph structure model or tree sequence specifically designed for a symmetric treatment of two responses in a regression setting. We assure the computational tractability of the model and a straightforward way of obtaining different conditional distributions. Using vine copulas the typical shortfalls of regression, as the need for transformations or interactions of covariates, collinearity or quantile crossings are avoided. We show a proof of concept by illustrating the copula-based bivariate quantiles for different copula distributions and by applying our model in a simulation study. Further, a data example emphasizes the benefits of bivariate modeling in contrast to two separate regressions for the two-response data set.

EG055 Room Virtual R27 CONTRIBUTIONS IN CAUSAL INFERENCE AND GRAPHICAL MODELS Chair: Dorota Kurowicka

E1076: **Regular vine copulas with strongly chordal pattern of (conditional) independence**
Presenter: **Dorota Kurowicka**, Delft University of Technology, Netherlands
Taking into account the (conditional) independence for a given data can simplify model estimation. A popular way of capturing the (conditional) independence is to use probabilistic graphical models. The relationship between strongly chordal graphs and m-saturated vines is proven. Moreover, an algorithm to construct a m-saturated vine structure corresponding to a strongly chordal graph is provided. This allows the (conditional) independence to be introduced into the regular vine copula model before its estimation. When the underlying data is sparse, the approach leads to a reduction of computational time and improves model estimation. Due to the reduction of model complexity, it is possible to evaluate all vine structures as well as to fit non-simplified vines. These advantages have been shown in the simulated and real data examples.

E1352: **Valid causal inference when using deep convolutional neural networks to control for highly structured covariates**
Presenter: **Mohammad Ghasempour**, Umea University, Sweden
Co-authors: Niloofar Moosavi, Xavier de Luna

Convolutional neural networks (CNN) have been successful in machine learning applications including image classification. When it comes to images, their success relies on their ability to consider the space invariant local features in the data. We consider the use of CNN to fit highly dimensional nuisance models in semiparametric estimation of a one-dimensional causal parameter: the average causal effect of a binary treatment. In this setting, nuisance models are functions of pre-treatment covariates that need to be controlled for. In an application where we want to estimate the effect of early retirement on a health outcome, we propose to use CNN to control for highly dimensional and time-structured covariates. Thus, CNN is used when fitting nuisance models explaining the treatment assignment and the outcome. These fits are then combined into an estimator having a nonparametric doubly robust property. Theoretically, we contribute by providing rates of convergence for CNN equipped with the rectified linear unit activation function and compare it to an existing result for feedforward neural networks. We also show when those rates guarantee uniformly valid inference for the proposed doubly robust estimator. A Monte Carlo study is provided where the performance of the proposed estimator is evaluated and compared with other strategies. Finally, we give results on a study of the effect of early retirement on later hospitalization using a database on the Swedish population.

E1115: **The dual PC algorithm for structure learning of Bayesian networks**
Presenter: **Enrico Giudice**, University of Basel, Switzerland
Co-authors: Jack Kuipers, Giusi Moffa

Learning the graphical structure of Bayesian Networks from observational data is a computationally challenging task and key to understanding data generating processes in complex applications. The Directed Acyclic Graph (DAG) of the Bayesian Network model is generally not identifiable from observational data, and a variety of methods exist to estimate the equivalence class of the DAG. Under certain assumptions, the popular PC algorithm can consistently recover the correct equivalence class by recursively testing for conditional independence (CI), starting from marginal independencies and progressively expanding the conditioning set. We propose the dual PC algorithm, a novel scheme to carry out the CI tests within the PC algorithm by leveraging the inverse relationship between covariance and precision matrices. Notably, the elements of the precision matrix coincide with partial correlations for Gaussian data. The algorithm then exploits block matrix inversions on the covariance and precision matrices to simultaneously perform tests on partial correlations of complimentary (or dual) conditioning sets. The multiple CI tests of the PC algorithm, therefore, proceed by first considering marginal and full-order CI relationships and progressively moving to central-order ones. Simulation studies indicate that the dual PC algorithm outperforms the classical PC algorithm both in terms of run time and in recovering the underlying network structure.

EG067 Room Virtual R35 CONTRIBUTIONS IN METHODOLOGICAL STATISTICS I Chair: Eugen Pircalabelu

E1611: **Estimation of treatment effects for multiple outcomes by using generalized linear models**
Presenter: **Shintaro Yuki**, Doshisha University, Japan
Co-authors: Hiroshi Yadohisa

A randomized controlled trial between two groups is considered. The objective is to identify a population with characteristics such that the test therapy is more effective than the control therapy. Such a population is called a subgroup. This identification can be made by estimating the treatment effect and identifying interactions between treatments and covariates. To date, many methods have been proposed to identify subgroups for a single outcome. There are also multiple outcomes, but they are difficult to interpret and cannot be applied to outcomes other than continuous values. We propose a multivariate regression method that introduces latent variables to estimate the treatment effect on multiple outcomes simultaneously. The proposed method introduces latent variables and adds Lasso sparsity constraints to the estimated loadings to facilitate the interpretation of the relationship between outcomes and covariates. The framework of the generalized linear model makes it applicable to various types of outcomes. Interpretation of subgroups is made by visualizing treatment effects and latent variables. This allows us to identify subgroups with characteristics that make the test therapy more effective for multiple outcomes. Simulation and real data examples demonstrate the effectiveness of the proposed method.

E1640: **Approximated variational inference based on data augmentation methods**
Presenter: **Cristian Castiglione**, University of Padova, Italy
Co-authors: Mauro Bernardi

Data augmentation is a powerful expedient that permits to formalize complicated statistical models through an equivalent convenient representation that relies on a set of auxiliary variables. This strategy is often employed for computational purposes to design iterative algorithms with closed-form updates, being fruitful for either optimization or simulation problems. Some remarkable examples are the augmented EM, the Gibbs sampling and the mean-field variational Bayes algorithms. Although their simplicity, data augmentation methods have been proved to suffer for low convergence rate and high sample autocorrelation, in the EM and Markov chain Monte Carlo cases, respectively. No theoretical results are available in the variational Bayes context, even though empirical experience suggests that different choices for the augmentation strategy can strongly affect the goodness of the posterior approximation. This gap is bridged by proving theoretically that the introduction of auxiliary variables leads to a systematic loss of information, which is measured as an increment of the Kullback-Liebler divergence between the approximated and the true posterior density, thereby reducing the global approximation accuracy. The validity of such a result is also supported by several data applications lying on logistic regression, quantile regression, and support vector machine classification.

E1485: **Wrapped Gaussian process functional regression for batch data on Riemannian manifolds**
Presenter: **Jinzhao Liu**, Newcastle University, United Kingdom
Co-authors: Jian Qing Shi

Regression is an essential and fundamental methodology in statistical analysis. Plenty of literature focuses on linear and nonlinear regression in the context of the Euclidean space. However, regression models in non-Euclidean spaces deserve more attention since people observed enormous manifold-valued data. Most existing regression models are nonviable in such a setup due to the lack of global vector space structure. Taking the advantage of massive manifold-valued data, we propose a concurrent functional regression model for batch data on a Riemannian manifold by estimating both mean structure and covariance structure simultaneously. The response variable is considered as a wrapped Gaussian process functional regression model. Nonlinear relationship between manifold-valued response variables and multiple Euclideancovariates can be captured by this model in which the covariates could be functional and scalar. The performance of our model has been tested on both generated data and real data, which endorses it as an effective and efficient tool in conducting functional data regression on a Riemannian manifold.

CO256 Room Virtual R31 ADVANCES IN TIME SERIES ECONOMETRICS Chair: Claudio Morana

C0176: **A new macro-financial condition index for the euro area**
Presenter: **Claudio Morana**, Universita di Milano Bicocca, Italy

A new time-domain decomposition is introduced for weakly stationary or trend stationary processes, based on trigonometric polynomial modeling of the underlying component of an economic time series. The method is explicitly devised to disentangle medium to long-term and short-term fluctuations in macroeconomic and financial series, to accurately measure the financial cycle and the concurrent long swings in economic activity.

The implementation of this decomposition is straightforward and relies on standard regression analysis and general to specific model reduction. Full support to the proposed method is provided by Monte Carlo simulation. We also provide a multivariate extension, involving sequential univariate decompositions and Principal Components Analysis. Based on this multivariate approach, we introduce a set of new composite indexes of macro-financial conditions for the euro area and assess their information content. In particular, concerning the current pandemic, the indicators suggest that most of the GDP contraction has been of short-term, cyclical nature. This is likely due to the prompt monetary and fiscal policy responses. Yet, our evidence suggests that the financial cycle might have currently achieved a peak area. Hence, the risk of further, deeper disruptions is high, particularly in so far as a new sovereign/corporate debt crisis were not eventually avoided.

C0187: **Uncertainty measures from partially rounded probabilistic forecast surveys**
Presenter: **Matthias Hartmann**, Deutsche Bundesbank, Germany
Co-authors: Alexander Glas

Although survey-based point predictions have been found to outperform successful forecasting models, corresponding variance forecasts are frequently diagnosed as heavily distorted. Forecasters who report inconspicuously low ex-ante variances often produce squared forecast errors that are much larger on average. We document the novel stylized fact that this variance misalignment is related to the rounding behavior of survey participants. Rounding may reflect that some survey participants employ a rather judgmental approach to forecasting as opposed to using a formal model. We use the distinct numerical accuracies of panelists' reported probabilities as a way to propose several alternatives and easily implementable corrections that 1. can be carried out in real-time, i.e., before outcomes are observed, and 2. deliver a significantly improved match between ex-ante and ex-post forecast uncertainty. According to our estimates, uncertainty about inflation, output growth and unemployment in the U.S. and the Euro area is higher after correcting for the rounding effect. The increase in the share of non-rounded responses in recent years also helps to understand the trajectory of survey-based average uncertainty during the years since the financial and sovereign debt crisis.

C0189: **Dimension reduction for high dimensional vector autoregressive models**
Presenter: **Gianluca Cubadda**, University of Rome Tor Vergata, Italy
Co-authors: Alain Hecq

The aim is to decompose a large dimensional vector autoregressive (VAR) model into two components, the first one being generated by a small-scale VAR and the second one being a white noise sequence. Hence, a reduced number of common factors generates the entire dynamics of the large system through a VAR structure. This modelling extends the common feature approach to high dimensional systems, and it differs from the dynamic factor model in which the idiosyncratic component can also embed a dynamic pattern. We show the conditions under which this decomposition exists. We provide statistical tools to detect its presence in the data and to estimate the parameters of the underlying small-scale VAR model. We evaluate the practical value of the proposed methodology by simulations as well as by an empirical application to a large set of US economic variables.

C1455: **Matrix inequality constraints for vector GARCH models with spillovers: A new (mixture) formulation**
Presenter: **Menelaos Karanasos**, Brunel University, United Kingdom
Co-authors: Yongdeng Xu, Alexandros Paraskevopoulos, Starvoula Yfanti

The purpose is to review and generalize results on the derivation of tractable non-negativity (necessary and sufficient) conditions for N-dimensional asymmetric power GARCH models. In practice, these constraints may not be fulfilled. To handle these cases we propose a new mixture formulation in order to eliminate some of these constraints. By using the exponential specification for some (but not all) of the conditional variables in the system we considerably reduce their dimensions. We also obtain new theoretical results about the second-moment structure and the optimal forecasts of such multivariate processes. An empirical example is included to show the effectiveness of the proposed method. We study the correlation dynamics among nineteen sectoral corporate bond indices. The time-varying cross-sector nexus is found to depend on economic fundamentals. The analysis further shows that the bond sectoral interconnectedness is highly vulnerable to crisis episodes, jeopardising the stability of the financial system.

CO888 Room Virtual R33 RECENT DEVELOPMENTS ON ECONOMETRICS: THEORY AND APPLICATIONS Chair: Ke Zhu

C1336: **Tests of unit root hypothesis with heavy-tailed heteroscedastic noises**
Presenter: **Rui She**, The Southwestern University of Finance and Economics, China

The unit-root testing with unspecified and heavy-tailed heteroscedastic noises is studied. A new weighted least squares estimation (WLSE) is designed to be used in the Dickey-Fuller (DF) test, of which the asymptotic normality is verified. However, the performance of the DF test strongly relies on the estimation accuracy of the asymptotic variance, which is not stable for dependent time series. To overcome this issue, we develop two novel unit-root tests by applying the empirical likelihood technique to the WLSE score equations. It is shown that both of the empirical likelihood-based tests converge weakly to a chi-squared distribution with one degree of freedom. Furthermore, the limiting theory is extended to the weighted M-estimation score equation. In contrast to existing unit-root tests for heavy-tailed time series, the empirical likelihood tests do not involve any estimators of the unknown parameters or any restrictions on the tail index of noise, which is of more practical appealing, and thus can be widely used in finance and econometrics. Extensive simulation studies are conducted to examine the effectiveness of the proposed methods.

C1407: **Quantiled moments by Cornish-Fisher expansion and its applications**
Presenter: **Ningning Zhang**, The University of Hong Kong, Hong Kong
Co-authors: Ke Zhu

The conditional moments play an important role in many financial applications. However, some parametric models for studying the conditional moments may exit model mis-specification problems and computation burden. To avoid these problems, a novel simple method is proposed to learn the conditional mean, variance, skewness, and kurtosis by using the classical Cornish-Fisher expansion. Our method provides an easy-to-implement non-parametric way to estimate the so-called quantiled moments, based on a sequence of estimated conditional quantiles. Some regression-based Wald tests are proposed to check the validity of our quantiled moments. Simulations show that the quantiled moments could be good proxies for their unobserved counterparts, and they exhibit robust performances across the choices of quantile estimation method and quantile level. As two important applications, the quantiled moments unveil unknown news impact functions and interactive effects among the conditional moments.

C1464: **Segmenting the time series via self-normalization**
Presenter: **Feiyu Jiang**, Fudan University, China
Co-authors: Xiaofeng Shao, Zifeng Zhao

A novel and unified framework is proposed for change-point estimation in multivariate time series. The method is fully nonparametric, enjoys effortless tuning and is robust to temporal dependence. Moreover, it treats change-point detection for a broad class of parameters (such as mean, variance, correlation and quantile) in a unified fashion. At the core of our method, we couple the self-normalization (SN) based tests with a novel nested local-window segmentation algorithm, which seems new in the growing literature of change-point analysis. Due to the presence of an inconsistent long-run variance estimator in the SN test, non-standard theoretical arguments are further developed to derive the consistency and convergence rate of the proposed SN-based change-point detection method. Extensive numerical experiments and relevant real data analysis are conducted to illustrate the effectiveness and broad applicability of the method in comparison with state-of-the-art approaches in the literature.

CO892 Room Virtual R34 ADDITIVE AND MULTIPLICATIVE TIME-VARYING GARCH MODELS Chair: Niklas Ahlgren

C0778: **Additive time-varying GARCH model**
Presenter: **Alexander Back**, Hanken School of Economics, Finland
Co-authors: Niklas Ahlgren, Timo Terasvirta

A GARCH model augmented by a time-varying intercept is proposed. The intercept is parameterized by a logistic transition function with rescaled time as the transition variable, which provides a flexible and simple way of capturing deterministic nonlinear changes in the conditional and unconditional variances. By making the intercept a smooth function of time, it is possible to capture changes that occur gradually, rather than abruptly as in regime-switching models. It is common for financial time series to exhibit these types of shifts. The time-varying intercept makes the model globally nonstationary but locally stationary. We use the theory of locally stationary processes to derive the asymptotic properties of the quasi maximum likelihood estimator (QMLE) of the parameters of the model. We show that the QMLE is consistent and asymptotically normally distributed. To corroborate the results, we provide a simulation study. An empirical application on stock returns of large US corporations demonstrates the usefulness of the model. We find that the persistence implied by the GARCH(1,1) parameter estimates is reduced by incorporating a time-varying intercept. Estimates of the GARCH parameters that suggest an integrated volatility model are diminished to lie within the stationary region when the ATV-GARCH model is fitted.

C0781: **Testing of parametric additive time-varying GARCH models**
Presenter: **Niklas Ahlgren**, Hanken School of Economics, Finland
Co-authors: Alexander Back, Timo Terasvirta

The aim is twofold. First, it develops a specification test for GARCH models with a time-varying intercept. The time-varying intercept is modelled by logistic transition functions with rescaled time as the transition variable. The model is an example of an additive decomposition of the conditional variance such that the conditional variance component is allowed to evolve smoothly over time. The model is called an additive time-varying (ATV-)GARCH model. The ATV-GARCH model is globally nonstationary but locally stationary. We derive Lagrange multiplier (LM) tests of GARCH against ATV-GARCH. The tests are based on auxiliary regressions. Despite the non-stationarity of the process, the LM statistics have standard asymptotic null distributions. The finite-sample properties of the tests are examined by simulations. Second, the article discusses a modelling strategy for ATV- and multiplicative time-varying (MTV-)GARCH models. The LM tests against ATV and MTV alternatives are not asymptotically independent and have power against each other. Both models accommodate deterministic changes in the amplitude of volatility clusters and the unconditional variance. The choice between these two types of models is an empirical question. A computational advantage of the additive model is that it is simpler to fit than the multiplicative model. The testing-based modelling strategy is illustrated by two empirical examples.

C0739: **Modelling non-stationarity robust variance interactions**
Presenter: **Cristina Amado**, University of Minho, Portugal

A multivariate generalisation of the multiplicative decomposition of the volatility is proposed within the class of conditional correlation GARCH models. The GARCH variance equations are multiplicatively decomposed into a deterministic non-stationary component describing the long-run movements in volatility and a short run dynamic component allowing for spillover effects between assets. The conditional correlations are assumed to be time-invariant in their simplest form or generalised into a flexible dynamic parametrisation. Parameters of the model are estimated equation-by-equation applying the maximisation by parts algorithm in the variance equations in the first step, and the correlation parameters estimated in the second step. An empirical application between four major spot exchange rates against the euro illustrates the usefulness of the model. Our results suggest that neglecting non-stationarity in the form of structural changes in the unconditional variance leads to spurious spillover effects. Furthermore, after modelling the variance equations accordingly, we also find evidence that some transmission mechanism of shocks persists which is supported by the presence of variance interactions robust to non-stationarity.

CG039 Room K0.18 (Hybrid 03) CONTRIBUTIONS IN HIGH-DIMENSIONAL ECONOMETRICS Chair: Guillaume Chevillon

C1423: **Modelling long memory with just one lag**
Presenter: **Guillaume Chevillon**, ESSEC Business School, France
Co-authors: Luc Bauwens, Sebastien Laurent

A large dimensional network or system can generate long memory in its components. Conditions have been derived under which the variables generated by an infinite-dimensional vector autoregressive model of order 1, a VAR(1), exhibit long memory. We go one step further and show how these asymptotic results can be put to practice for finite sample modelling and inference regarding series with long-range dependence that belongs to a network or a large system. We propose to use a VAR(1), or an AR(1)-X when the VAR(1) model is estimated equation by equation, whose parameters we shrink to generic conditions matching previous work. The proposal significantly outperforms ARFIMA and HAR models when forecasting a nonparametric estimate of the log of the integrated variance (i.e., log(MedRV)) of 250 assets, the annual productivity growth recorded in 100 industrial sectors in the U.S., as well as seasonally adjusted historic monthly streamflow series recorded in 97 localisations of the Columbia river basin.

C1085: **High dimensional generalised least squares**
Presenter: **Aikaterini Chrysikou**, Kings College, University of London, United Kingdom
Co-authors: George Kapetanios, Ilias Chronopoulos

Inference for high dimensional linear models with serially correlated errors is developed. We examine the latter using the Lasso under the assumption of strong mixing in the covariates and error process, allowing for fatter tails in their distribution. While the Lasso estimator performs poorly under such circumstances, we estimate via penalised FGLS the parameters of interest and extend the asymptotic properties of the Lasso under more general conditions. The theoretical results indicate that the non-asymptotic bounds for stationary dependent processes are sharper, while the rate of the Lasso under general conditions appears slower as $T, p \to \infty$. Further, we use the de-biased Lasso to perform inference on the parameters of interest. Using simulated data, we find that with the debiased generalised least squares estimator, the t-tests appear more powerful and correctly sized, while the true value of the parameter is included in the 95% confidence interval with satisfying coverage rates at different levels of parameter sparsity.

C1558: **Pooling dynamic conditional correlation models**
Presenter: **Bram van Os**, Econometric Institute, Erasmus University Rotterdam, Netherlands
Co-authors: Dick van Dijk

The Dynamic Conditional Correlation (DCC) model has become an extremely popular tool for modeling the time-varying dependence of asset returns. However, applications to large cross-sections have been found to be problematic, due to the curse of dimensionality. We propose a novel DCC model with Conditional LInear Pooling (CLIP-DCC) which endogenously determines an optimal degree of commonality in the correlation innovations, allowing a part of the update to be of reduced dimension. In contrast to existing approaches such as the Dynamic EquiCOrrelation (DECO) model, the CLIP-DCC model does not restrict long-run behavior, thereby naturally complementing target correlation matrix shrinkage approaches. Empirical findings suggest substantial benefits for a minimum-variance investor in real-time. Combining the CLIP-DCC model with target shrinkage yields the largest improvements, confirming that they address distinct parts of uncertainty of the conditional correlation matrix.

CG035 Room K0.19 (Hybrid 04) CONTRIBUTIONS IN FINANCIAL NETWORKS Chair: Jozef Barunik

C0192: **Good contagion: What do networks say about policy transmission**
Presenter: **Kumushoy Abduraimova**, Durham University, United Kingdom
Contagion is frequently considered as something bad, as a small initial shock amplifying into a systemic crisis, as financial distress propagating from one bank to another, or as a spread of infectious disease. A different perspective is taken. The focus is on how contagion can facilitate monetary policy transmission throughout the network. We develop a multi-layer network-based contagion centrality measure and apply it to analyse the transmission efficiency of the European Central Banks interest rate policy measures undertaken to tackle the low for long inflation. This is the first study, to our knowledge, that addresses policy transmission from a network perspective. The main finding of this analysis is that the policy transmits most efficiently during severe bearish contagion and is least efficient during intense bullish contagion. This finding could be attributed to the level of attention that markets pay to policy announcements during turmoil and calm periods.

C1582: **Transitory and persistent networks**
Presenter: **Jozef Barunik**, UTIA AV CR vvi, Czech Republic
Co-authors: Michael Ellington
A novel framework is proposed to measure connectedness from variance decompositions. The approach accounts for the characteristics of shocks creating network structures. Using frequency domain techniques, we measure connectedness that forms on the transitory and persistent component of shocks. We outline a procedure to test for statistical differences in transitory and persistent network connectedness. Monte Carlo evidence shows our measures reliably track connectedness and correctly identify statistical differences. We show that our connectedness measures enhance our understanding of systemic risks emerging from sectoral uncertainty networks. Therefore, they may serve as a monitoring tool for macro-prudential supervisors and investors alike.

C1656: **A macroprudential view on post-trade risk reduction services**
Presenter: **Yuliang Zhang**, LSE, United Kingdom
Co-authors: Luitgard Veraart
The consequences of post-trade risk reduction services for systemic risk in derivatives markets are analyzed. The focus is on portfolio rebalancing, which is a mechanism of injecting new trades to reduce the overall counterparty exposure, and portfolio compression, which is a mechanism to reduce the outstanding notional amount by trades termination and replacement. We first provide a mathematical characterisation of (optimal) portfolio rebalancing. Then, we explore the effects of these services on the financial system from a network perspective by considering contagion arising from only partial repayments in networks of variation margin payments. We provide sufficient conditions for portfolio rebalancing to reduce systemic risk. We also investigate the effects under a scenario where financial institutions react to stress strategically and make delayed payments.

CG029 Room Virtual R30 CONTRIBUTIONS IN CAUSALITY Chair: Michael Knaus

C1729: **Doubly robust estimation of conditional average treatment effect for the treated under DID framework**
Presenter: **XingLei Deng**, Xiamen University, China
A new estimator is proposed to estimate the conditional average treatment effect for the treated(CATT) under the difference in difference(DID) framework. Comparing to the ordinary estimator, our new estimator not only doubly robust, but also more stable and more efficient under some situation. Simulation shows that the new estimator has good finite sample performance. Finally, we use this new estimator to estimate the mean treatment effect of increasing minimum wage on counties' unemployment rates, we find that counties with different median income levels are affected differently by the policy.

C1690: **Inferences for partially conditional quantile treatment effect model**
Presenter: **Shengfang Tang**, Wang Yanan Institute for Studies in Economics, China
A new model, termed the partially conditional quantile treatment effect (PCQTE) model, is proposed to characterize the heterogeneity of treatment effect conditional on some predetermined variable(s). We show that the partially conditional quantile treatment effect is identified under the assumption of selection on observables, which leads to a semiparametric estimation procedure in two steps: first, parametric estimation of the propensity score function and then, nonparametric estimation of conditional quantile treatment effect. Under some regularity conditions, the consistency and asymptotic normality of the proposed semiparametric estimator are derived. More importantly, a specification test is seminally proposed in quantile regression literature, to test whether there exists heterogeneity for PCQTE across sub-populations based on the Cramer-von Mises type criterion. The asymptotic properties of the proposed test statistic are investigated, including consistency and asymptotic normality. Finally, the performance of the proposed methods is illustrated through Monte Carlo experiments and an empirical application on estimating the effect of the first-time mothers smoking during pregnancy on the baby's birth weight conditional on mothers age and testing whether the partially conditional quantile treatment effect varies across different mothers age.

C1675: **Statistical causality: What more can we learn about our data**
Presenter: **AB Zaremba**, University College London, Quantitative Risk Solutions Lab, United Kingdom
Co-authors: Gareth Peters
A novel testing framework for statistical causality has been developed in general classes of multivariate nonlinear time series models. Our framework allows us to study causality in the trend, volatility, or both, and accommodates a range of structural features that are important when modelling financial time series, including long memory. However, we want to emphasise the distinction between choosing an exact model and finding a causal relation. With our framework, we test the dependence with regards to the structures that are specified, and we show a range of examples of good performance even for misspecified models. We then focus on the example of commodity futures data and show the usefulness of testing for causality under different model specifications as a way to explore the data and the potentially complex dependence relationships. We point out what can be learned from comparing statistical causality tests based on GPs to analogous tests based on linear regression, explain what makes the latter overconfident, and show a breakdown of steps that can adjust for the overconfidence.

CG027 Room Virtual R32 CONTRIBUTIONS IN ECONOMETRIC MODELLING Chair: Roberto Casarin

C1631: **Score-driven generalized Poisson model**
Presenter: **Giulia Carallo**, Ca' Foscari University of Venice, Italy
Co-authors: Roberto Casarin, Dario Palumbo
A new score driven model for integer data is introduced. In particular, we introduce a dynamic conditional score model where the series has Generalized Poisson conditional distribution (GP-DCS), for the location and scale parameters. We provide a Bayesian inference framework and an efficient posterior approximation procedure based on Markov Chain Monte Carlo. An application to fire data shows that the proposed DCS model is well suited for capturing persistence in the conditional moments and in the over-dispersion feature of the data.

C0532: **Estimating quantile treatment effects for panel data**
Presenter: **Mingfeng Zhan**, Xiamen University, China
Co-authors: Cai Zongwu, Ying Fang, Ming Lin

A factor-based model has been previously proposed to estimate the average treatment effect with panel data. A quantile treatment effect model for panel data is now proposed, to characterize the distributional effect of a treatment. We utilize the relationship between conditional cumulative distributional function (CDF) and unconditional CDF to estimate the counterfactual quantile for the treated unit. Also, we derive the asymptotic properties for the proposed quantile treatment effect estimator, together with discussing the choice of control units and covariates. A simulation study is conducted to illustrate our method. Finally, the proposed method is applied to estimate the quantile treatment effects of introducing CSI 300 index futures trading on both the log-return and volatility of the stock market in China.

C1672: **The limits of measurement**

Presenter: **Ioannis Paraskevopoulos**, Universidad Pontificia Comillas, Spain

The focus is on the dependence of the unknown stochastic solution on the known stochastic chain in Banach space. We extend the limits of computation beyond Hilbert space, as Banach space is somewhere in between the Hilbert and another space endowed with a measure but without a norm. In particular, we examine whether the evolution process depends on its known history. We argue that multidimensional integration has its dual in reflexive Banach space if the solution obeys the functional central limit theorem and its error is between the limits of dilation and erosion. Hence there is an isomorphism for closed and semi-open intervals and the desired inner product to define the orthogonality conditions for the open interval $[0,1)$. One possible scenario will be, as previously suggested, no model exists and reversibility will be unattainable as it would map to infinite possible initial points of the past. In this case all strong solutions will be of this type, $S_k = \left(\prod_{-\infty}^{k} \xi_j\right) V$. Where V is an independent process. We derive the orthogonality conditions and show that the unknown stochastic unique solution would depend on the observable history of the sequence, as $S_m = \prod_{-\infty}^{m} \rho_j S_{m-1}$.

Saturday 18.12.2021 11:10 - 12:50 Parallel Session C – CFE-CMStatistics

EO240 Room K0.18 (Hybrid 03) TRACK: BAYESIAN SEMI- AND NON-PARAMETRIC METHODS I Chair: Antonio Canale

E0518: **Distributed inference for generalized linear models with variable selection**
Presenter: **Luca Maestrini**, University of Technology Sydney, Australia
Co-authors: Matias Quiroz, Feng Li

A computationally efficient framework is proposed for fitting generalized linear models and performing variable selection through a distributed system. The dataset is partitioned into numerous subsets and maximum likelihood estimation is performed on each subset. The resulting estimates are combined together to form a pseudo-likelihood approximation to the likelihood function which can be linked to a penalization. The final parameter estimation is conducted via a fast approximate inference method. We illustrate our results on prominent generalized linear models using different priors for variable selection.

E0726: **Spike-and-slab variational bayes for high-dimensional survival analysis**
Presenter: **Michael Komodromos**, Imperial College London, United Kingdom
Co-authors: Sarah Filippi, Marina Evangelou, Kolyan Ray

In recent years variational Bayes (VB) has presented itself as a viable alternative to MCMC, particularly in situations where scalability is key. We follow such developments and present a VB approximation to sparse high-dimensional Bayesian proportional hazards models. Within our VB approximation we utilise a mean-field spike-and-slab variational family, thereby offering mechanisms for variable selection, coefficient estimation and uncertainty quantification. We demonstrate the performance in a variety of simulation settings, as well as demonstrate applicability to real-world datasets.

E0664: **Generalized infinite factorization**
Presenter: **Lorenzo Schiavon**, University of Padova, Italy
Co-authors: Antonio Canale, David Dunson

Factorization models express a statistical object of interest in terms of a collection of simpler objects. For example, a matrix or tensor can be expressed as a sum of rank-one components. However, in practice, it can be challenging to infer the relative impact of the different components as well as the number of components. A popular idea is to include infinitely many components having an impact decreasing with the component index. The motivation comes from two limitations of existing methods: (1) lack of careful consideration of the within component sparsity structure; and (2) no accommodation for grouped variables and other non-exchangeable structure. We propose a general class of infinite factorization models that address these limitations. Theoretical support is provided, practical gains are shown in simulation studies, and an ecology application focusing on modeling bird species occurrence is discussed.

E1681: **Learning-augmented count-min sketches via Bayesian nonparametrics**
Presenter: **Stefano Favaro**, University of Torino and Collegio Carlo Alberto, Italy
Co-authors: Emanuele Dolera, Stefano Peluchetti

The count-min sketch (CMS) is a randomized data structure that provides estimates of tokens frequencies in a large data stream using a compressed representation of the data by random hashing. We present a Bayesian nonparametric (BNP) approach to CMS, and then develop a novel learning-augmented CMS under power-law data streams. We assume that tokens in the stream are drawn from an unknown discrete distribution, which is endowed with a Pitman-Yor process (PYP) prior. By means of distributional properties of the PYP, we compute the posterior distribution of a tokens frequency in the stream, given the hashed data, and in turn, corresponding BNP estimates. Applications to synthetic and real data show that our approach achieves a remarkable performance in the estimation of low-frequency tokens. This is known to be a desirable feature in the context of natural language processing, where it is indeed common in the context of the power-law behaviour of the data.

EO060 Room K0.19 (Hybrid 04) CLUSTERING OF COMPLEX DATA STRUCTURES Chair: Maria Brigida Ferraro

E0254: **Spatial weighted robust clustering of multivariate time series with an application to COVID-19 pandemic**
Presenter: **Angel Lopez Oriona**, Universidade da CoruÃśa, Spain
Co-authors: Jose Vilar, Pierpaolo Durso

A fuzzy clustering model for multivariate time series based on the quantile cross-spectral density and principal component analysis is improved. The extension consists of (i) a weighting system that assigns a weight to each principal component in accordance with its importance concerning the underlying clustering structure and (ii) a penalization term allowing to take into account the spatial information. The iterative solutions of the new model, which employs the exponential distance in order to gain robustness against outlying series, are derived. A simulation study shows that the introduction of the weighting system substantially enhances the effectiveness of the former approach. The behaviour of the extended model in terms of the spatial penalization term is also analysed. An application involving multivariate time series of mobility indicators concerning COVID-19 pandemic highlights the usefulness of the proposed technique.

E0523: **Fuzzy spectral clustering for document data sets**
Presenter: **Irene Cozzolino**, Universita La Sapienza, Italy
Co-authors: Maria Brigida Ferraro, Peter Winker

In recent years, spectral clustering methods have been successfully applied in the field of text classification. The success of these methods is largely based on their solid theoretical foundations which do not make any assumption on the global structure of the data. Despite their good performance in text classification, little has been done in the field of clustering. In this regard, a crucial point for every spectral clustering algorithm is the construction of a similarity matrix to use as input of the algorithm, which should well describe the intrinsic nature of the data. To enhance the clustering performance, and motivated by the inherent sequential nature of text data, a new similarity measure is introduced, which is obtained as a weighted combination of sequence and set similarities. Indeed, the only use of sequence similarities ignores the non-sequential part which might be similar in content too. Moreover, we introduce a novel fuzzy version of spectral clustering for text data to use in combination with the proposed similarity matrix. The adequacy of the new document clustering method is evaluated by means of benchmark and real data sets.

E0655: **A novel bi-clustering algorithm for Hilbert data**
Presenter: **Agostino Torti**, Politecnico di Milano, Italy
Co-authors: marta galvani, Alessandra Menafoglio, Piercesare Secchi, Simone Vantini

The problem of bi-clustering for the analysis of Hilbert data is considered with the aim of simultaneously clustering the rows and columns of a data matrix whose entries are objects for which a meaningful Hilbert space structure can be identified. A definition of ideal bi-cluster for Hilbert data is given and a novel bi-clustering algorithm - called HC2 (i.e, Hilbert Cheng and Church) - is developed. The HC2 relies on a non-parametric deterministic iterative procedure capable of finding bi-clusters in a data matrix where each cell contains an object, possibly belonging to a multidimensional space. The introduced algorithm is very flexible and allows one to discover different types of bi-clusters depending on the model chosen to define the concept of ideal bi-cluster for the problem at hand. Simulation studies are performed to show the potentials of the

introduced method. The HC2 algorithm is finally applied to the analysis of the regional railway service in the Lombardy region with the aim of identifying recurrent patterns in the passengers' daily access to trains and/or stations, thus supporting correct management of the service.

E0667: **Clustering random intervals: A new approach based on a similarity measure**
Presenter: **Ana Belen Ramos-Guajardo**, University of Oviedo, Spain

A hierarchical method for clustering random intervals is proposed. The idea is to group random intervals based on the similarity of their corresponding expected values. In this way, two random intervals will be joined if the degree of similarity of their expected values can be assumed to be greater than or equal to a certain degree. Such a similarity degree between each pair of random intervals can be analyzed by means of a two-sample similarity bootstrap test providing finally a p-value matrix. Thus, the higher the p-value obtained, the greater the similarity between both random intervals. The iterative clustering algorithm suggested comprises an objective stopping criterion that leads to statistically similar clusters that are different from each other. Lastly, a comparative simulation study and an application of the method to a real case are shown.

EO438 Room K0.20 (Hybrid 05) RECENT DEVELOPMENTS IN EXTREME VALUE THEORY AND METHODS Chair: Gilles Stupfler

E0186: **Estimation of the cure rate for distributions in the Gumbel maximum domain of attraction under insufficient follow-up**
Presenter: **Mikael Escobar-Bach**, Universita d'Angers, France
Co-authors: Ross Maller, Ingrid Van Keilegom, Muzhi Zhao

Estimating the cured proportion from survival data which may include observations on cured subjects, that is, those who never experience the event of interest, is a critical task in practice. Any proposed estimator can only be expected to perform well when the follow-up period is sufficient, in some sense. We propose an adjustment that ameliorates the problem when follow-up is insufficient and under the assumption that the survival distribution of those susceptible to the event belongs to the Gumbel maximum domain of attraction, since many commonly used lifetime distributions have this property. We use extrapolation techniques from extreme value theory to derive a non-parametric estimator of the cure proportion, which is consistent and approximately normally distributed under certain assumptions, and performs well in simulation studies. We illustrate with an application to survival data where patients with different stages of breast cancer have varying degrees of follow-up.

E0753: **Extreme conditional expectile estimation in heavy-tailed heteroscedastic regression models**
Presenter: **Antoine Usseglio-Carleve**, Avignon Universita, France
Co-authors: Gilles Stupfler, Stephane Girard

Expectiles define a least-squares analogue of quantiles. They have been the focus of a substantial quantity of research in the context of actuarial and financial risk assessment over the last decade. The behaviour and estimation of unconditional extreme expectiles using independent and identically distributed heavy-tailed observations has been investigated recently. We build a general theory for the estimation of extreme conditional expectiles in heteroscedastic regression models with heavy-tailed noise; our approach is supported by general results of independent interest on,residual-based extreme value estimators in heavy-tailed regression models, and is intended to cope with covariates having a large but fixed dimension. We demonstrate how our results can be applied to a wide class of important examples, among which linear models, single-index models as well as ARMA and GARCH time series models. The estimators are showcased on a numerical simulation study and on real sets of actuarial and financial data.

E1037: **Extremal dependence between maxima of concomitants**
Presenter: **Amir Khorrami Chokami**, University of Turin, Italy
Co-authors: Marie Kratz

The problem of finding methods to describe the extremal dependence among multiple time series has rapidly become attractive in recent years, due to the vast variety of fields where its practical implications are of interest. However, providing handy tools to assess such dependence is still challenging. An empirical method has been recently developed by Dacorogna and Cadena, where the authors provide a statistical approach to explore the dependence among extreme risks. The purpose is to develop further this problem from a theoretical point of view. The mathematical formalization that we propose involves the concept of concomitants of order statistics, widely studied in the literature. We focus on the asymptotic dependence between maxima of concomitants: Specifically, order a bivariate sequence of n i.i.d. random variables (X,Y) on the basis of the X-variable, and call the extreme set of the sequence (X_n,Y_n) the subset of couples where the first component is one of the k largest order statistics (k fixed). Consider the vector formed by the maxima of the concomitants belonging to the extreme set and to its complementary set. We study how the bivariate extremal dependence of (X,Y) influences the asymptotic joint distribution of the two maxima of concomitants. Revisiting a pivotal work, we propose an alternative way to tackle the problem, which allows us to consider the cases where upper tail dependence is present.

E1395: **Extreme expectile estimation from short-tailed data**
Presenter: **Boutheina Nemouchi**, ENSAI, France
Co-authors: Boutheina NEMOUCHI, Abdelaati Daouia, Gilles Stupfler

Expectile estimation has been recently investigated extensively from the perspective of extreme value theory. Attention has been, however, restricted to heavy-tailed distributions. We focus on the less-discussed problem of estimating extreme expectiles from short-tailed data. We present extrapolated expectile estimators based on extreme quantiles and develop their asymptotic theory. Then, we provide Monte Carlo evidence of their superiority relative to competing estimators extrapolating intermediate asymmetric least squares expectiles. Finally, we discuss concrete examples of application.

EO378 Room Virtual R18 RECENT ADVANCES IN FDA Chair: Siegfried Hoermann

E0767: **Functional data analysis and censoring**
Presenter: **David Kraus**, Masaryk University, Czech Republic
Co-authors: Stanislav Nagy

Functional data analysis is often complicated by the fact that the information collected about the curves is distorted due to incomplete, fragmentary, discrete, noisy or otherwise imperfect observation. We focus on censoring. We discuss the estimation of characteristics of the distribution of functional data and the prediction of functional observations.

E0826: **Sieve bootstrap memory parameter in long-range dependent stationary functional time series**
Presenter: **Han Lin Shang**, Macquarie University, Australia

A sieve bootstrap procedure is applied to quantify the estimation uncertainty of long-memory parameters in stationary functional time series. To estimate the long-memory parameter, we use a semiparametric local Whittle estimator, where discrete Fourier transform and periodogram are constructed from the first set of principal component scores, via a functional principal component analysis. The sieve bootstrap procedure uses a general vector autoregressive representation of the estimated principal component scores. It generates bootstrap replicates that adequately mimic the dependence structure of the underlying stationary process. For each bootstrap replicate, we first compute the estimated first set of principal component scores and then apply the semiparametric local Whittle estimator to estimate the memory parameter. By taking quantiles of the estimated memory parameters from these bootstrap replicates, we can construct confidence intervals of the long-memory parameter. As measured by coverage probability differences between the empirical and nominal coverage probabilities at three levels of significance, we demonstrate the advantage of using the sieve bootstrap in comparison to the asymptotic confidence intervals based on normality.

E0897: **Pivotal tests for relevant differences in the second order dynamics of functional time series**
Presenter: **Anne van Delft**, Columbia University, United States
Co-authors: Holger Dette

Motivated by the need to statistically quantify differences between modern (complex) datasets which commonly result as high-resolution measurements of stochastic processes varying over a continuum, novel testing procedures are proposed to detect relevant differences between the second-order dynamics of two functional time series. In order to take the between-function dynamics into account that characterize this type of functional data, a frequency domain approach is taken. Test statistics are developed to compare differences in the spectral density operators and in the primary modes of variation as encoded in the associated eigenelements. Under mild moment conditions, we show convergence of the underlying statistics to Brownian motions and construct pivotal test statistics. The latter is essential because the nuisance parameters can be unwieldy and their robust estimation infeasible, especially if the two functional time series are dependent. Besides these novel features, the properties of the tests are robust to any choice of frequency band enabling also to compare energy contents at a single frequency. The finite sample performance of the tests are verified through a simulation study and are illustrated with an application to fMRI data.

E1018: **Non-linear function-on-function regression via neural nets**
Presenter: **Matthew Reimherr**, Pennsylvania State University, United States
Co-authors: Aniruddha Rao

A new class of non-linear function-on-function regression models for functional data using neural networks is introduced. We propose a framework using a hidden layer consisting of continuous neurons, called a continuous hidden layer, for functional response modeling and give two model-fitting strategies, Functional Direct Neural Networks (FDNN) and Functional Basis Neural Networks (FBNN). Both are designed explicitly to exploit the structure inherent in functional data and capture the complex relations existing between the functional predictors and the functional response. We fit these models by deriving functional gradients and implementing regularization techniques for more parsimonious results. We demonstrate the power and flexibility of our proposed method in handling complex functional models through extensive simulation studies as well as real data examples.

EO066 Room Virtual R20 THE STEIN METHOD AND STATISTICS Chair: Robert Gaunt

E0570: **About Stein kernels**
Presenter: **Yvik Swan**, Universite libre de Bruxelles, Belgium

Stein kernels (a.k.a. Stein covariance kernels) are important functionals associated with probability distributions that have recently attracted quite a bit of attention due to their role e.g. in controlling the speed of convergence in some multivariate central limit theorems. We will provide a brief history of these quantities, and review some of their most important applications.

E0716: **Parameter estimators for non-normalized statistical models based on Stein characterizations**
Presenter: **Steffen Betsch**, Karlsruhe Institute of Technology, Germany
Co-authors: Bruno Ebner, Bernhard Klar

A new estimation method for the parameters of non-normalized models consisting of smooth density functions on the real line is presented. For everyone to be on the same page, a short summary of the issues with non-normalized densities is given. The eventual estimation procedure is based on a recently introduced characterization for the respective probability distributions that is derived from a classical identity which underlies some applications of Stein's method. These characterizations are stated and used to construct minimum distance parameter estimators. The new method is compared, in simulations, with different approaches from the machine learning literature that tackle the same problem. Moreover, connections to inferential methods based on Stein discrepancies are discussed.

E0723: **On testing randomness of binary images**
Presenter: **Bruno Ebner**, Karlsruhe Institute of Technology, Germany

New methods are presented for testing the randomness of binary image data. The tests are based on so-called Minkowski functionals such as the area, the perimeter or the Euler-Poincar characteristic. We derive the limit null distribution of the test statistics by means of Stein's method in the Kolmogorov distance using dependency graphs. A Monte Carlo simulation study shows that the tests are able to detect alternatives and we apply it to data related to some irrational numbers and mathematical constants such as pi, e, square root of 2, Euler-Mascheroni constant and the Catalan constant.

E0845: **Approximations for random sums with equally correlated summands**
Presenter: **Fraser Daly**, Heriot-Watt University, United Kingdom

Let $Y = X_1 + \cdots + X_N$ be a sum of a random number of random variables, where the random variable N is independent of the X_j. Such random sums arise in many applications, including in the areas of financial risk, hypothesis testing and physics. Classically, the X_j are assumed to be independent, in which case central limit theorems and other distributional approximation results for Y are well known. However, this assumption of independent X_j is unrealistic in many applications. We relax this restriction, instead of assuming that these random variables come from a generalized multinomial model. In this setting, we prove error bounds in Gaussian and Poisson approximations for Y which allow us to investigate the effect of the correlation parameter on the quality of the approximation, while also providing competitive bounds in the special case of independent X_j. We also derive error bounds for Gamma approximation in the special case where N has a Poisson distribution. The proofs make use of Stein's method in conjunction with size-biased and zero-biased couplings.

EO314 Room Virtual R21 RANDOM MATRIX THEORY AND RELATED FIELDS Chair: Bowen Gang

E0425: **Phase transition of eigenvector distribution for spiked models**
Presenter: **Zhigang Bao**, Hong Kong University of Science and Technology, Hong Kong

Recent work on the eigenvector distribution of the spiked random matrix models will be reviewed. We will mainly focus on the supercritical and critical regimes of the Baik-Ben Arous-Peche (BBP) phase transition. Together with the previous result of Bloemendal-Knowles-Yin-Yau on the subcritical regime, these results fully depict the phase transition of the eigenvectors of the spiked models.

E0516: **CLT for linear spectral statistics of large dimensional kendall's rank correlation matrices and its applications**
Presenter: **Zeng Li**, Southern University of Science and Technology, China

The focus is on the limiting spectral behaviors of large dimensional Kendall's rank correlation matrices generated by samples with independent and continuous components. The statistical setting covers a wide range of highly skewed and heavy-tailed distributions since we do not require the components to be identically distributed, and do not need any moment conditions. We establish the central limit theorem (CLT) for the linear spectral statistics (LSS) of Kendall's rank correlation matrices under the Marchenko-Pastur asymptotic regime, in which the dimension diverges to infinity proportionally with the sample size. We further propose three nonparametric procedures for high dimensional independent test and their limiting null distributions are derived by implementing this CLT. Our numerical comparisons demonstrate the robustness and superiority of our proposed test statistics under various mixed and heavy-tailed cases.

E0661: **On limiting spectral distribution of Spearman's and Kendall's rank correlation matrices with general dependence**
Presenter: **Cheng Wang**, Shanghai Jiao Tong University, China

The spectral distribution of eigenvalues of Spearman's and Kendall's rank correlation matrices is studied, under the assumption that the observations are i.i.d. random vectors with non-paranormal distributions. It is the first result on rank correlation matrices with dependence.

E0676: **SIMPLE: Statistical Inference on Membership Profiles in Large Networks**
Presenter: **Xiao Han**, University of Science and Technology of China, China

Network data is prevalent in many contemporary big data applications in which the common interest is to unveil important latent links between different pairs of nodes. Yet a simple fundamental question of how to precisely quantify the statistical uncertainty associated with the identification of latent links still remains largely unexplored. We propose the method of statistical inference on membership profiles in large networks (SIMPLE) in the setting of a degree-corrected mixed membership model, where the null hypothesis assumes that the pair of nodes share the same profile of community memberships. Under some mild regularity conditions, we establish the exact limiting distributions of the two forms of SIMPLE test statistics under the null hypothesis and contiguous alternative hypothesis. They are the chi-square distributions and the non-central chi-square distributions, respectively, with degrees of freedom depending on whether the degrees are corrected or not. We also address the important issue of estimating the unknown number of communities and establishing the asymptotic properties of the associated test statistics. The advantages and practical utility of our new procedures in terms of both size and power are demonstrated through several simulation examples and real network applications.

EO154 Room Virtual R22 MODELING SPATIOTEMPORAL DATA Chair: Mattias Villani

E0519: **Robust real-time delay predictions in a network of high-frequency urban buses**
Presenter: **Hector Rodriguez-Deniz**, Linkoping University, Sweden
Co-authors: Mattias Villani

Providing transport users and operators with accurate forecasts on travel times is challenging due to a highly stochastic traffic environment. We develop a robust model for real-time bus travel time prediction that depart from Gaussian assumptions by using Student-t errors. The proposed approach uses spatiotemporal characteristics from the route and previous bus trips to model short-term effects, and date/time variables and Gaussian processes for long-run forecasts. The model allows for flexible modeling of mean, variance and kurtosis spaces. We propose algorithms for Bayesian inference and for computing probabilistic forecast distributions. Experiments are performed using data from high-frequency buses in Stockholm, Sweden. Results show that Student-t models outperform Gaussian ones in terms of log-posterior predictive power to forecast bus delays at specific stops, which reveals the importance of accounting for predictive uncertainty in model selection. Estimated Student-t regressions capture typical temporal variability between within-day hours and different weekdays. Strong spatiotemporal effects are detected for incoming buses from immediately previous stops, which is in line with many recently developed models. We finally show how Bayesian inference naturally allows for predictive uncertainty quantification, e.g. by returning the predictive probability that the delay of an incoming bus exceeds a given threshold.

E1163: **Matern fields on graphs and their edges**
Presenter: **Jonas Wallin**, Lund University, Sweden

The focus is on Gaussian processes defined on edges of a Graph, which for example is relevant when modelling traffic data on road networks. It is difficult to define a proper covariance function on such a topology, and even harder to define a covariance function with Markov properties, which is one of our main goals. We show how one can construct Matern-type processes using precision operators acting on each edge independently. Then through (random) boundary conditions, such as Neumann-Kirchhoff conditions, one can tie the edges together to form a process acting on the entire graph. For models of this type, we demonstrate how to efficiently evaluate log-likelihoods and to perform Kriging prediction.

E1238: **Prehospital resource optimization**
Presenter: **Patrik Ryden**, Umea University, Sweden

Prehospital care in Sweden has about 660 ambulances, respond to about 1.2 million emergency calls per year, and costs more than 0.5 billion euro per year. An aging population, urbanization and medical progress demand flexible prehospital care. The goal is to develop processes and tools that make it possible to organize ambulance units and operations in an optimal way. Based on big and complex alarm-data (we have detailed data on all alarms during the last 5 years), advanced statistical modelling and large-scale data-driven simulations we have developed tools to compare allocations (how the ambulances are placed and scheduled) under user-defined future scenarios. The solution makes it easy to highlight the implications for specific regions and patient groups. An important and challenging part of this modelling is the spatio-temporal modelling of the alarms. The alarm intensity varies over the day, between weekdays and there is also a clear seasonal trend. In addition, the spatial distribution of the alarms is highly concentrated on the road network. We will present the problem, the data and some early solutions. Furthermore, evaluation of spatio-temporal models will be discussed.

E1536: **Spatial-temporal stochastic fields with embedded dynamics**
Presenter: **Krzysztof Podgorski**, Lund University, Sweden

General non-stationary spatial-temporal surfaces that involve dynamics governed by velocity fields are discussed. In that approach, the dynamics are introduced by embedding deterministic velocities into a stochastic spatial-temporal Gaussian model. In this way, a dynamically inactive stochastic field with a given spatial and temporal covariance structure gains dynamics that in general follow a deterministic pattern. We make an important connection between the resulting stochastic field and underlying deterministic dynamics by demonstrating that in the case of isotropic spatial dependencies, the observed random velocities are centered at the velocities of the underlying physical flow. Additionally, we discuss strategies for simulation of such fields and give a foundation for statistical fitting and prediction procedures that are based on the obtained results.

EO276 Room Virtual R23 RECENT DEVELOPMENTS IN YUIMA PACKAGE AND RELATED TOPICS Chair: Nakahiro Yoshida

E0508: **yuima.PPR: New developments for the point process in the YUIMA package.**
Presenter: **Lorenzo Mercuri**, University of Milan, Italy

The purpose is to present and discuss yuima classes and methods that allow the user to simulate and estimate a Point Process Regression Model (PPR). The PPR model can be seen as a generalization of a self-exciting point process, since it is possible to consider external covariates that explain the behaviour of the intensity process. To manage a PPR model, two new objects have been introduced. The first object belongs to yuima.PPR-class, contains the mathematical structure of a PPR model and eventually the dataset. The last object belongs to the yuima.PPR.qmle-class and is filled with a dataset, the model description and the estimated parameters obtained from the considered dataset. We discuss also how to use these objects to manage the CARMA-Hawkes process recently proposed in actuarial science to model the insurance claims. The CARMA-Hawkes process is a generalization of the standard Hawkes where the exponential kernel is substituted by the CARMA kernel. The main advantage of this model is its ability to reproduce a more complex dependence structure compared with the autocovariance generated by the Hawkes process.

E0766: **Asymptotic expansion formulas for diffusion processes based on the perturbation method**
Presenter: **Emanuele Guidotti**, University of Neuchatel, Switzerland
Co-authors: Nakahiro Yoshida

Diffusion processes are a class of models that plays a prominent role in describing the time-continuous evolution of phenomena in the natural and social sciences. However, only in very few cases, the stochastic differential equation driving the process can be analytically solved. Based on

the perturbation method, we present asymptotic expansion formulas to generate accurate approximations to the solution of arbitrary diffusions. In particular, we expand the characteristic function of the process. Then, the approximated expectation and moments are computed by differentiation, and the approximated transition density is written in terms of Hermite polynomials by applying Fourier transform. The computational efficiency, accuracy, and flexibility of the method are assessed via experiments conducted against closed-form solutions and Monte Carlo simulations in tasks involving density approximation, expectations, moments, filtering, and functionals of generic diffusion processes.

E0851: **Expanding Levy-SDE related functions in YUIMA package**
Presenter: **Hiroki Masuda**, Kyushu University, Japan

Some recent studies are presented on statistical inference for the driving noise of an ergodic Markovian stochastic differential equation (SDE), from the viewpoints of both theory and implementation in R-package YUIMA. The process is supposed to be observed at a high frequency over long-time period. After a brief overview of the Gaussian quasi-likelihood inference, we will present newly expanded Levy-SDE related functions in YUIMA. In particular, we will illustrate some integrated use of the internal functionalities for the quasi-maximum likelihood estimation (qmle) and for handling noise non-Gaussianity (yuima.law-class).

E1642: **Porting the Yuima package to Python: Difficulties and feasibility**
Presenter: **Francesco Iafrate**, University of Rome La Sapienza, Italy

According to the latest charts Python is ranked as the most popular programming language in the world, as well as the most used by the Data Science community. Currently, however, it appears that an extensive and unified framework for simulation and estimation of SDEs is not available for Python. The extension of the Yuima package to the Python user base can open the doors for a wide variety of new applications and it can bring forth new stimuli for the Yuima community. Porting Yuima to the new platform poses several technical challenges caused by the inherently different features between R and Python. We investigate such difficulties and propose a tentative construction of a new package resembling the structure of Yuima that is suitable for the Python programming style.

EO298 Room Virtual R24 CHALLENGES AND OPPORTUNITIES IN ANALYSING CLINICAL DATA Chair: Eleni-Rosalina Andrinopoulou

E1029: **Optimal design of clinical trials when outcome values are missing**
Presenter: **Robin Mitra**, Cardiff University, United Kingdom
Co-authors: Stefanie Biedermann

The presence of missing values complicates analyses. When designing experiments, such as clinical trials, missing values are particularly problematic when constructing optimal designs, as it is not known which values are missing at the design stage. When information about the missing data mechanism is available, it is possible to incorporate this information into the optimality criterion that is used to find designs. The areas of missing data and optimal design are briefly reviewed. Some of the specific challenges are then considered when finding optimal designs for the analysis of clinical trials when outcome values could be missing.

E0917: **Bayesian methods for handling missing values in complex settings**
Presenter: **Nicole Erler**, Erasmus Medical Center, Netherlands

Missing values are a challenge regularly encountered in the analysis of real-world data. Multiple imputation via a fully conditional specification (FCS) is popular to address the missing data problem. The separation of imputation and analysis and the specification of a set of univariate full-conditional models are attractive features that allow re-using the imputed data and facilitate a straightforward specification of imputation models for mixed-type variables. In multi-level data, time-to-event analyses or settings involving non-linear associations, however, important assumptions of the FCS approach are likely violated, leading to biased results. A fully Bayesian analysis, in which the parameters of interest are estimated jointly with the missing values, is an attractive alternative in such complex settings. The joint distribution of response variable(s), incomplete variables and parameters can be conveniently split into a sequence of (univariate) conditional distributions, allowing the choice of appropriate distributions for mixed-type variables while not requiring the inclusion of the response into a linear predictor. This makes the Bayesian approach suitable for highly complex substantive models, such as multivariate joint models of longitudinal and survival data. Moreover, any complex association structures specified in the substantive model(s) are automatically taken into account during imputation, ensuring compatibility between all sub-models involved.

E0801: **Predictive modeling approaches to personalized medicine: A comparison of regression-based methods**
Presenter: **David van Klaveren**, Erasmus MC University Medical Center, Netherlands

The benefits and harms of medical treatments vary substantially between individual patients. Predictive modeling approaches to personalized medicine are designed to predict the benefit of one treatment over another for an individual patient. A regression model including many interactions between treatment and patient risk factors may be an obvious choice for predicting treatment benefit. However, simulated data will be used to show that including fewer treatment interactions often leads to better treatment benefit predictions. This will be further illustrated with the recently proposed Syntax Score II (SSII)-2020 which was developed to predict the difference in 10-year mortality when treating complex coronary artery disease patients with heart bypass surgery rather than coronary stenting. Cox regression was first used in the SYNTAX trial data ($n = 1,800$) to develop a prognostic index (PI) for mortality over a 10-year horizon consisting of 7 clinical predictors of mortality. Second, a Cox model was fitted which included the treatment, the PI and pre-specified treatment interactions with type of disease and with anatomical disease complexity. In contrast to its more flexible predecessor SSII-2013, SSII-2020 was well calibrated for treatment benefit at 10 years post-procedure, both at cross-validation in the same data and at external validation in new data.

E0346: **Co-data weighted elastic net**
Presenter: **Mark van de Wiel**, Amsterdam University Medical Centers, Netherlands
Co-authors: Mirrelijn van Nee

Co-data stands for co-mplementary data. As opposed to covariates it contains information on the variables rather than on the samples. E.g. presence of a gene in a particular pathway, or a p-value that quantifies association strength between gene and outcome in a related, external study. Small sample size is a common challenge for clinical high-dimensional studies, e.g. due to cost restrictions or rarity of the disease. We show that the use of co-data can alleviate this challenge to some extent when the aim is to develop predictive signatures. From a methodological perspective, the focus lies on empirical Bayes techniques to incorporate the co-data information in the predictor, e.g. by adapting multiple penalty weights in the elastic net setting. Efficient computation of those penalties is a computational hurdle, which we take by an approximation that is of fairly general use. Ideas and software are illustrated by cancer genomics examples.

EO718 Room Virtual R25 RECENT DEVELOPMENTS IN CAUSAL INFERENCE Chair: BaoLuo Sun

E0499: **On proximal causal inference with synthetic controls**
Presenter: **Xu Shi**, University of Michigan, United States

The focus is on evaluating the impact of an intervention when time series data on a single treated unit and multiple untreated units are observed in pre- and post- treatment periods. A synthetic control (SC) method was previously proposed as an approach to relax the parallel trend assumption in difference-in-differences methods. The term SC refers to a weighted average of control units built to match the treated unit's pre-treatment outcome trajectory, such that the SC's post-treatment outcome predicts the treated unit's unobserved potential outcome under no treatment. The treatment

effect is then estimated as the difference in post-treatment outcomes between the treated unit and the SC. A common practice to estimate the weights is to regress the pre-treatment outcomes of the treated unit on that of the control units using ordinary or weighted least squares. However, it has been established that these estimators can fail to be consistent. We introduce a proximal causal inference framework for the SC approach and formalize identification and inference for both the SC weights and the treatment effect on the treated unit. We further extend the traditional linear model to nonlinear models allowing for binary and count outcomes which are currently under-studied in SC literature. We illustrate our proposed methods with simulation studies and an application to evaluate the 1990 German Reunification.

E0638: **MRCIP: A robust Mendelian randomization method accounting for correlated and idiosyncratic pleiotropy**
Presenter: **Zhonghua Liu**, The University of Hong Kong, Hong Kong

Mendelian randomization (MR) is a powerful instrumental variable (IV) method for estimating the causal effect of an exposure on an outcome of interest even in the presence of unmeasured confounding by using genetic variants as IVs. However, the correlated and idiosyncratic pleiotropy phenomena in the human genome will lead to biased estimation of causal effects if they are not properly accounted for. We develop a novel MR approach named MRCIP to account for correlated and idiosyncratic pleiotropy simultaneously. We first propose a random-effect model to explicitly model the correlated pleiotropy and then propose a novel weighting scheme to handle the presence of idiosyncratic pleiotropy. The model parameters are estimated by maximizing a weighted likelihood function with our proposed PRW-EM algorithm. Moreover, we can also estimate the degree of the correlated pleiotropy and perform a likelihood ratio test for its presence. Extensive simulation studies show that the proposed MRCIP has improved performance over competing methods. We also illustrate the usefulness of MRCIP on two real datasets. The R package for MRCIP is publicly available at https://github.com/siqixu/MRCIP.

E0824: **The variety of no direct effect assumptions in dynamic treatment regimes**
Presenter: **Lin Liu**, Shanghai Jiao Tong University, China

Recent theoretical results on semiparametric efficiency theory when some direct effects are known to be absent based on background knowledge are presented. Then we will showcase how such results can be applied to precision medicine or personalized decision making to more efficiently utilize the data. In particular, we will compare and contrast several works that leverage such "no direct effect" assumptions, and show that we can provide a unified picture of all these works. If time permitted, we will also present our adventure on how neural networks can be used to estimate parameters efficiently in this context, i.e. under the "no direct effect" assumption.

E0901: **Identifying effects of multiple treatments in the presence of unmeasured confounding**
Presenter: **Wang Miao**, Peking University, China
Co-authors: Wenjie Hu, Elizabeth Ogburn, Xiaohua Zhou

Identification of treatment effects in the presence of unmeasured confounding is a persistent problem in the social, biological, and medical sciences. The problem of unmeasured confounding in settings with multiple treatments is most common in statistical genetics and bioinformatics settings. Recently there have been a number of attempts to bridge the gap between these statistical approaches and causal inference, but these attempts have either been shown to be flawed or have relied on fully parametric assumptions. We propose two strategies for identifying and estimating causal effects of multiple treatments in the presence of unmeasured confounding. The auxiliary variables approach leverages variables that are not causally associated with the outcome; in the case of a univariate confounder, our method only requires one auxiliary variable, unlike existing instrumental variable methods that would require as many instruments as there are treatments. An alternative null treatments approach relies on the assumption that at least half of the confounded treatments have no causal effect on the outcome, but does not require a priori knowledge of which treatments are null. The identification strategies do not impose parametric assumptions on the outcome model and do not rest on the estimation of the confounder.

EO770 Room Virtual R26 CAUSAL MEDIATION ANALYSIS Chair: Anita Lindmark

E0511: **Causal mediation analysis: From simple to more robust estimation strategies**
Presenter: **Trang Nguyen**, Johns Hopkins Bloomberg School of Public Health, United States

The aim is to provide practitioners of causal mediation analysis with a better understanding of estimation options. We take as inputs two familiar strategies (weighting and regression-based prediction) and a simple way of combining them (weighted models) and show how we can generate a range of estimators with different modeling requirements and robustness properties. The primary goal is to help build an intuitive appreciation for robust estimation that is conducive to sound practice. A second goal is to provide a menu of estimators that practitioners can choose from for the estimation of marginal natural (in)direct effects. The estimators generated from this exercise include some that coincide or are similar to existing estimators and others that have not previously appeared in the literature. We note several different ways to estimate the weights for cross-world weighting based on three expressions of the weighting function, including one that is novel; and show how to check the resulting covariate and mediator balance. We use a random continuous weights bootstrap to obtain confidence intervals, and also derive general asymptotic (sandwich) variance formulas for the estimators. The estimators are illustrated using data from an adolescent alcohol use prevention study.

E0411: **Semiparametric estimation for causal mediation analysis with multiple causally ordered mediators**
Presenter: **Xiang Zhou**, Harvard University, United States

Causal mediation analysis concerns the pathways through which treatment affects an outcome. While most of the mediation literature focuses on settings with a single mediator, a flourishing line of research has examined settings involving multiple mediators, under which path-specific effects (PSEs) are often of interest. We consider estimation of PSEs when the treatment effect operates through K causally ordered, possibly multivariate mediators. In this setting, the PSEs for many causal paths are not nonparametrically identified, and we focus on a set of PSEs that are identified under Pearl's nonparametric structural equation model. These PSEs are defined as contrasts between the expectations of 2^{K+1} potential outcomes and identified via what we call the generalized mediation functional (GMF). We introduce an array of regression-imputation, weighting, and hybrid estimators, and, in particular, two $K+2$-robust and locally semiparametric efficient estimators for the GMF. The latter estimators are well suited to the use of data-adaptive methods for estimating their nuisance functions. We establish the rate conditions required of the nuisance functions for semiparametric efficiency. We also discuss how our framework applies to several estimands that may be of particular interest in empirical applications. The proposed estimators are illustrated with a simulation study and an empirical example.

E0558: **Simulating hypothetical interventions on multiple mediators: Extending methods and practical guidance**
Presenter: **Margarita Moreno-Betancur**, University of Melbourne and Murdoch Children's Research Institute, Australia
Co-authors: John B Carlin

Many research questions concern the pathways presumed to mediate an association, particularly in epidemiological research. Invariably, the translational intent of such research is to inform potential intervention targets, but until recently mediation effect definitions did not acknowledge this interventional intent. Recent work proposed a novel framework conceptualising mediation effects by mapping to a hypothetical "target" randomised trial evaluating mediator interventions. This approach is particularly relevant for mediators that do not correspond to well-defined interventions, which arise frequently, and perhaps can only be addressed by considering hypothetical interventions that would shift the mediators' distributions. The approach proposes specifying a target trial to capture the research question in the form of (a measure of) the impact of shifting joint mediator distributions to user-specified distributions. These estimand assumptions are distinguished from identifiability assumptions, which are needed to emulate the effects with the observed data. By its nature, the approach is context-specific. Drawing on learnings from applications to several longitudinal cohort studies, some of which are already published, further developments of the method are presented with a focus on

alternative approaches to effect definition according to the question in diverse contexts. A workflow that will assist researchers in applying the method in practice is described.

E0196: **Nonlinear mediation analysis with high-dimensional mediators whose causal structure is unknown**
Presenter: **Wen Wei Loh**, Ghent University, Belgium
Co-authors: Beatrijs Moerkerke, Tom Loeys, Stijn Vansteelandt

With multiple possible mediators on the causal pathway from treatment to an outcome, the focus is on the problem of decomposing the effects along multiple possible causal paths through each distinct mediator. Fine-grained decompositions under Pearl's path-specific effects framework necessitate stringent assumptions, such as correctly specifying the causal structure among the mediators, and no unobserved confounding among the mediators. In contrast, interventional direct and indirect effects for multiple mediators can be identified under much weaker conditions. However, current estimation approaches require (correctly) specifying a model for the joint mediator distribution, which can be difficult when there is a high-dimensional set of possibly continuous and non-continuous mediators. We avoid the need to model this distribution, by developing a definition of interventional effects previously suggested for longitudinal mediation. We propose a novel estimation strategy that uses non-parametric estimates of the (counterfactual) mediator distributions. Non-continuous outcomes can be accommodated using non-linear outcome models. Estimation proceeds via Monte Carlo integration. The procedure is used to assess the effect of a microRNA expression on the three-month mortality of brain cancer patients via expression values of multiple genes.

EO094 Room Virtual R27 ADVANCES IN MULTIVARIATE FUNCTIONAL DATA ANALYSIS Chair: Sophie Dabo

E0693: **Investigating spatial scan statistics for multivariate functional data**
Presenter: **Camille Frevent**, University of Lille, France
Co-authors: Mohamed Salem Ahmed, Sophie Dabo, Michael Genin

In some research fields, such as environmental surveillance, pollution sensors are deployed in a geographical area. In a context where these sensors measure simultaneously the concentrations of many pollutants at regular intervals over a long period of time, environmental experts may search for environmental black spots, that can be defined as geographical areas characterized by elevated concentrations of pollutants. To this end, the development of spatial scan statistics for multivariate functional data indexed in space is very relevant. The proposed methods are derivated from a MANOVA test statistic for functional data, an adaptation of the Hotelling T^2-test statistic, and a multivariate extension of the Wilcoxon rank-sum test statistic. The performances of the methods are investigated through a simulation study and they are applied on multivariate functional data to search for spatial clusters of abnormal daily concentrations of air pollutants in the north of France in May and June 2020.

E0734: **Adaptive optimal estimation of irregular mean and covariance functions**
Presenter: **Steven Golovkine**, Crest, France
Co-authors: Nicolas Klutchnikoff, Valentin Patilea

Straightforward nonparametric estimators for the mean and the covariance functions of functional data are proposed. The random trajectories are, not necessarily differentiable, have unknown regularity, and are measured with error at discrete design points. The measurement error could be heteroscedastic. The design points could be either randomly drawn or common for all curves. The definition of the nonparametric estimators depends on the local regularity of the stochastic process generating the data. First, a simple estimator of this local regularity which takes strength from the replication and regularization features of functional data is given. Next, the mean and the covariance functions are estimated using the "smoothing first, then estimate" approach. The new estimators achieve optimal rates of convergence. They can be applied with both sparsely or densely sampled curves, are easy to calculate and update, and perform well in simulations. Simulations built upon a real data example on power consumption illustrate the effectiveness of this approach.

E0793: **Some consistent results for the kernel-based Bayes classifier for functional data with values in a submanifolds**
Presenter: **Anne Francoise Yao**, Universite Clermont Auvergne/LMBP, France
Co-authors: Pamela Llop, Chafik Samir, Papa Mbaye

The problem of classification for functional data which lies in a finite-dimensional submanifold of a Hilbert space H is considered. We discuss the choice of the kernel which is a central issue in this setting and study the consistency of the corresponding kernel-based Bayes classifier based on n independent copies (X_i, Y_i) of the couple of variables (X, Y) where X belongs to a d submanifold of H. We motivate the practical interest of such a classifier through some applications. Basically, because we deal with functional data, the classification problem requires achieving several issues as finding a suitable representation. Then, through the applications, we illustrate the behavior of the Bayes classifier of interest, based on several representations, including the L_2 representation as well as spherical (normalization of the functional data) representation.

E1531: **cfda R package for categorical functional data**
Presenter: **Cristian Preda**, University of Lille, France
Co-authors: Vincent Vandewalle, Quentin Grimonprez

Data represented by paths of a continuous-time stochastic jump process will be considered in the framework of functional data analysis. Dimension reduction, clustering and regression will be illustrated throughout the cfda R package on simulated and real data applications.

EO152 Room Virtual R33 TOPICS ON HIGH-DIMENSIONAL METHODS Chair: Eugen Pircalabelu

E0668: **Node clustering in large-scale graphical models**
Presenter: **Andreas Alfons**, Erasmus University Rotterdam, Netherlands
Co-authors: Daniel Touw, Ines Wilms

Graphical models can represent conditional dependency structures among a large number of variables in a compact manner, especially when relying on edge-sparsity as a simplifying structure. Since data are nowadays measured and analyzed at ever-higher resolutions or more disaggregate levels, similarity in the conditional dependency structure among different nodes becomes a new, interesting guiding principle for dimensionality reduction. We develop an unsupervised node-clustering method for large-scale graphical models. Our aim is to find clusters of nodes in the network that share their conditional dependency structure. To this end, we leverage recent advances from convex clustering within a regularization framework, and we show how node-clustering can be efficiently combined with other popular penalization structures for graphical models such as edge-sparsity. We compare the methods estimation and clustering performance to several other state-of-the-art benchmark methods.

E0777: **Sparse selection and prediction with high-dimensional categorical data**
Presenter: **Wojciech Rejchel**, University of Warsaw, Poland
Co-authors: Piotr Pokarowski, Szymon Nowakowski

Sparse prediction with categorical data is challenging even for a moderate number of variables, because one parameter is roughly needed to encode one category or level. The group lasso is a well known and efficient algorithm for the selection of continuous or categorical variables, but all estimates related to a selected factor usually differ, so a fitted model may not be sparse. To make a group lasso solution sparse, we propose to merge levels of the selected factor, if a difference between its corresponding estimates is less than some predetermined threshold. We prove that under weak conditions our algorithm recovers the true, sparse linear or logistic model even for the high-dimensional scenario, that is when a number of parameters is greater than a learning sample size. To our knowledge, selection consistency has been proven many times for different algorithms

fitting sparse models with categorical variables, but our result is the first for the high-dimensional scenario. Numerical experiments show the satisfactory performance of the method.

E1226: **Robust approaches for sufficient dimension reduction**
Presenter: **Andreas Artemiou**, Cardiff University, United Kingdom

Sufficient Dimension Reduction (SDR) is a supervised dimension reduction framework to address dimension reduction mainly in regression and classification settings. Classic methods in SDR involve the use of inverse moments. We will discuss new methods to address the outliers in SDR methodology by using multivariate medians.

E1290: **A stable network inference procedure for high dimensional data**
Presenter: **Emilie Devijver**, CNRS, France
Co-authors: Remi Molinier, Melina Gallopin

The stability of variable selection procedures is crucial on high dimensional data: one hopes that active variables do not depend on the observed sample, in the sense that new observations would not change the active set. In the context of network inference based on Gaussian Graphical Models, l1-penalized log-likelihood methods are not stable when the number of observations is limited. Adding a structure to the estimation problem can lead to more stable results. We demonstrate theoretical guarantees for the stability of network inference based on hierarchical clustering and a non-asymptotic model selection criterion. Unlike state-of-the-art methods to deal with stability, the inference procedure is not based on data resampling. The theoretical guarantees are derived from the stability properties of single-linkage hierarchical clustering, based on topological considerations. The proposed network inference method is particularly relevant in real data analysis when a large number of observations is difficult to obtain, such as network inference from gene expression data. Numerical experiments, on simulated and real datasets, support the theoretical guarantees.

EO890 Room Virtual R36 MIXTURE MODELLING Chair: Sollie Millard

E1016: **Feature selection in mixture of logistic regression models using the modified elastic-net penalty**
Presenter: **Salomi Millard**, University of Pretoria, South Africa
Co-authors: Sollie Millard, Mohammad Arashi, Frans Kanfer, Gaonyalelwe Maribe

Datasets with a relatively large number of highly correlated features are often found in applications of finite mixture regression models. Furthermore, the contribution of each feature towards the response variable differs in the respective components of the mixture model. This creates a complex feature selection problem. Penalised regression methods are frequently used to perform feature selection whilst addressing the issues that arise due to multicollinearity. We consider the use of the novel modified elastic-net (MEnet) penalty for statistical analysis and feature selection in a finite mixture of logistic regressions settings. An extensive simulation study is performed to demonstrate the properties pertaining to feature selection and classification accuracy of this approach. The methodology is also applied to real-world examples.

E0787: **Parsimonious mixture of multivariate mean-mixture of normal distributions**
Presenter: **Mehrdad Naderi**, University of Pretoria, South Africa
Co-authors: Andriette Bekker

Heterogeneity occurs in various problems of multivariate analysis where the data comes from different latent groups. The finite mixture (FM) model is a model-based statistical tool commonly exploited to identify these unobserved groups. Despite the widespread use and attractive properties of the normal distribution, practical researchers believe that the traditional normally-based FM (FM-N) model might not achieve robust inference when asymmetric features exist in data. A mixture of multivariate mean-mixture of normal (FM-MMN) distributions is postulated to address this potential issue. In addition to the parameters of the FM-N model, the proposed FM model has two vector/scaler parameters, in each component, for controlling skewness and mild heavy tails. Maximum likelihood parameter estimates are carried out by implementing an expectation-maximization (EM) type algorithm. The parsimony version of the FM-MMN distributions is also introduced by employing an eigenvalue decomposition of the component covariance matrices. Finally, the utility of the proposed methodology is illustrated by conducting a simulation study and analyzing real data examples.

E1011: **A possible solution to the label-switching problem in fitting nonparametric mixture of regressions**
Presenter: **Sphiwe Skhosana**, University of Pretoria, South Africa
Co-authors: Frans Kanfer, Sollie Millard

A nonparametric mixture of regressions (NMR), where the component regression functions have an unknown functional form but are assumed to be smooth functions of the covariate, provide a flexible approach to the analysis of heterogeneous regression relationships. The nonparametric regression functions are typically estimated over a set of grid points using local likelihood estimation via the Expectation-Maximization (EM) algorithm. However, maximizing each local likelihood function does not guarantee that the component labels will match at each grid point. We, therefore, have a potential label-switching problem. We propose a possible solution to the label switching problem which relies on the smoothness assumption of the nonparametric regression functions. Since the component labels at each grid point contain similar information about the data, we simultaneously maximize the local likelihood functions for each set of component labels and consider, as our final model, the solution that results in the smoothest component regression functions.

E0965: **Contributions to mixture GLMs with application in sequenced data**
Presenter: **Michelle de Klerk**, University of Pretoria, South Africa
Co-authors: Frans Kanfer, Sollie Millard

Mixture generalised linear models (MGLMs) is an effective statistical method in the study of sequenced data as it considers all genes and identifies latent clusters within the data. The semi-parametric mixture of generalised linear models (SPMGLM) applied on differentially expressed gene data is addressed. The SPMGLM considers a non-parametric specification of the link function which gives access to a larger subset of distributions in the exponential family. This yields a more flexible modelling structure used in identifying differentially expressed genes compared to traditional parametric modelling solutions. The performance of this technique is demonstrated via a simulation study.

EO501 Room Virtual R37 PROJECTION PURSUIT I Chair: Nicola Loperfido

E0512: **Some generalizations of the Henze theorem**
Presenter: **Haruhiko Ogasawara**, Otaru University of Commerce, Japan

The Henze theorem gives the decomposition of the skew-normally (SN) distributed variable in truncated and untruncated independent normal variables. The pseudo normal (PN) distribution was introduced by the author using sectional truncation as an extension of the SN and the closed skew-normal (CSN), which use only single hidden or latent truncation. Under sectional truncation, it is shown that a similar decomposition holds in the PN, where the moment generating function is used for the derivation. The derivation is also the third proof of the Henze theorem, which was originally derived by analytical and probabilistic methods. It is shown that a similar decomposition holds in the cases of the normal finite/infinite mixture with the normal density weights (NN distributions).

E0964: **Clustering via a semi-parametric density estimation**
Presenter: **Mahdi Salehi**, University of Pretoria, Iran
Co-authors: Andriette Bekker, Mohammad Arashi

The idea behind density-based clustering is to associate groups with the connected components of the level sets of the density of the data to be estimated by a nonparametric method. This approach claims some advantages over both distance- and model-based clustering. Some researchers developed this technique by proposing a graph theory-based method for identifying local modes of the underlying density being estimated by the well-known kernel density estimation (KDE) with normal and t kernels. The aim is to improve the performance of the density-based clustering by using a semi-parametric KDE with a more flexible family of kernels belonging to skew-symmetric distribution. Finding optimum bandwidth under the mentioned kernels is another main result where we shrink the bandwidth more than the one obtained under the normality assumption. Finally, some illustrative examples follow.

E1397: **A weighted k-medoids algorithm for clustering time series' projections**
Presenter: **Raffaele Mattera**, University of Naples Federico II, Italy
Co-authors: Germana Scepi

Time-series clustering is one of the most common techniques used to discover similar structures in a dataset with dynamic objects. The main issue in time series clustering lies in the computation of a proper distance. A lot of approaches, based on statistical model parameters or on time series features, have been proposed in the literature. Some clustering approaches do not consider as units the single time series but their projections. In this case, it is very important to define a peculiar distance, taking into account the characteristics of the observations. In this framework, we propose a kMEDOID-type algorithm based on an optimal weighting scheme for multiple distances. The weights, obtained by minimizing the weighted squared distance between each i-th object from its k-th medoid, reflect the importance of the information contained in each distance. The performance of the proposed fast algorithm will be evaluated by comparing the results with those obtained using well-known clustering approaches. Furthermore, an application to real datasets is provided.

E1413: **Entropy weighted model-based clustering of skewed and heavy tailed time series**
Presenter: **Massimiliano Giacalone**, University of Naples - Federico II, Italy
Co-authors: Raffaele Mattera, Karina Gibert

The goal of clustering is to identify common structures in a data set by forming groups of homogeneous objects. The observed characteristics of many economic time series motivated the development of classes of distributions that can accommodate properties such as heavy tails and skewness. Thanks to its flexibility, the Skewed Exponential Power Distribution (also called Skewed Generalized Error Distribution) ensures a unified and general framework for clustering possibly skewed and heavy-tailed time series. A clustering procedure of model-based type is developed, assuming that the time series are generated by the same underlying probability distribution but with different parameters. Moreover, we propose to optimally combine the estimated parameters to form the clusters with an entropy weighing k-means approach. Moreover, by applying skewness or kurtosis-based projection pursuit, the resulting interesting projections can be used as the input of the clustering procedure with a different distributional assumption. The usefulness of the proposal is showed by means of application to financial time series.

EO603 Room Virtual R39 EXPERIMENTAL DESIGN Chair: Heiko Grossmann

E0652: **Optimal two-level designs under model uncertainty**
Presenter: **Steven Gilmour**, KCL, United Kingdom
Co-authors: Pi-Wen Tsai

Two-level designs are widely used for screening experiments where the goal is to identify a few active factors which have major effects. Most work on two-level designs is based on the effect hierarchy assumption that lower-order effects are of more importance than higher-order effects, so the focus is on two-level designs with level balance and pairwise orthogonality. We apply the model-robust Q_B criterion for the selection of optimal two-level designs by incorporating experimenters' prior knowledge on the importance of each effect into the optimality criterion. We find a smooth relationship between the choice of designs and the experimenters' prior beliefs. Additionally, we provide a coordinate exchange algorithm for the construction of Q_B-optimal designs without the restrictions of level-balance and pairwise orthogonality.

E0936: **J-characteristics for 3-level response surface designs**
Presenter: **Jose Nunez Ares**, KU Leuven, Belgium
Co-authors: Peter Goos

The quality of two-level orthogonal screening designs is often evaluated by means of J-characteristics. We extend this work to three-level response surface designs, define the J-characteristics of an arbitrary 3-level response surface design, and provide a simple expression that relates the J-characteristics with the design points. An attractive feature of the J-characteristics is that they do not only quantify the degree of aliasing between effects, but also describe the morphology of a design, namely, the pattern of zeros. We use the J-characteristics to revisit some existing results for Orthogonally Minimally Aliased Response Surface Designs. Finally, we define a set of J-characteristics for 3-level response surface designs tailored to the case where a second-order effects model is of interest. Using these tailored J-characteristics, we propose an expression for the resolution and the aberration of 3-level response surface designs and we illustrate its usefulness with a set of 3,246 7-factor RSDs extracted from literature.

E0427: **Finding the optimal proportions in constrained mixture experiments**
Presenter: **Stefanie Biedermann**, The Open University, United Kingdom
Co-authors: Shroug Alzahrani

Experiments involving mixtures are conducted in a variety of areas, for example, in food processing or chemical research. The experimental region is constrained naturally, as the proportions of all ingredients must sum to one. Additional constraints may arise when there are bounds on the proportions; for example, a cake must contain a minimum percentage of flour to have the right texture and flavour. Often, interest is in finding the optimal mixture, e.g. the proportions of flour, butter, sugar and eggs that will result in the nicest cake. We propose new modelling approaches for mixture experiments and compare these with standard models from the literature. We then investigate the benefits of optimal designs for these models and discuss various open problems in the area.

E0928: **Optimal discrete choice designs via hypergraphs**
Presenter: **Frank Roettger**, Universita de Geneve, Switzerland
Co-authors: Rainer Schwabe, Thomas Kahle

Multinomial discrete choice models are studied to compare m unstructured alternatives in choice sets of constant size $k < m$. To describe the choice sets, we use a representation by uniform hypergraphs, where each choice set forms a hyperedge on k vertices in the complete hypergraph. In this model, finding D-optimal designs corresponds to a parameterized convex optimization problem. This relates by the Kiefer–Wolfowitz equivalence theorem to directional derivatives being non-positive in general and zero for choice sets in the support of the design. We show that via the linear Farris transform of the inverse information matrix, the system of directional derivatives is determined from a hyperedge-edge incidence matrix. For designs supported on $\binom{m}{2}$ choice sets the number of equations in the system of directional derivatives equals the number of unknowns, such that the system has a unique solution when the corresponding submatrix of the hyperedge-edge incidence matrix is invertible. In this case, the design weights are easily obtained from the inverse information matrix. In the Bradley–Terry paired comparison model ($k = 2$), the edge-edge incidence matrix is diagonal, and therefore always invertible. This implies that for designs supported on all $\binom{m}{2}$ edges, we always obtain the optimal design weights via a simple matrix inversion. Furthermore, this allows us to derive optimal weights for any design that is supported on a decomposable

graph.

EO178 Room Virtual R40 SPORT ANALYTICS Chair: Christophe Ley

E1263: **Multi-task Gaussian processes models for functional data and application to the prediction of swimming performances**
Presenter: **Arthur Leroy**, University of Sheffield, United Kingdom

Gaussian process regression is a common tool of supervised learning that provides a convenient probabilistic framework, leading to predictions with proper uncertainty quantification. However, the learning procedure in such models generally focuses on hyper-parameters estimation of the covariance structure rather than the prior mean of the process. Therefore, prediction quality might severely decrease with an inappropriate prior mean as we move away from observation points. A multi-task extension of the GP framework is introduced, where data are supposed to come from several individuals sharing some structure altogether. This approach offers more reliable predictions even when a new individual is observed on a few or sparse input locations. Then, the model is enhanced with a clustering component to provide cluster-specific GP predictions. We handle talent identification in sports, and illustrate the approach with this application involving performance swimming datasets. We will see how the proposed algorithm provides reliable probabilistic predictions of future performances while simultaneously allocating swimmers into clusters of similar individuals.

E1376: **Bayesian models for prediction of the set-difference in volleyball**
Presenter: **Ioannis Ntzoufras**, AUEB, Greece
Co-authors: Vasilios Palaskas, Sotirios Drikos

The aim is to study and develop Bayesian models for the analysis of volleyball match outcomes as recorded by the set-difference. Due to the peculiarity of the outcome variable (set-difference) which takes discrete values from -3 to 3, we cannot consider standard models based on the usual Poisson or binomial assumptions used for other sports such as football/soccer. Hence, the first and foremost challenge was to build models appropriate for the set-differences of each volleyball match. We consider two major approaches: a) an ordered multinomial logistic regression model and b) a model based on a truncated version of the Skellam distribution. For the first model, we consider the set-difference as an ordinal response variable within the framework of multinomial logistic regression models. Concerning the second model, we adjust the Skellam distribution in order to account for the volleyball rules. We fit and compare both models with the standard vanilla structure commonly used in team sports modelling. Both models are fitted, illustrated and compared within a Bayesian framework using data from both the regular season and the play-offs of the season 2016/17 of the Greek national men's volleyball league A1.

E1401: **Introducing regularisation to generalised joint regression modelling and its application to football and sports**
Presenter: **Hendrik van der Wurp**, TU Dortmund University, Germany
Co-authors: Andreas Groll

When modelling the bivariate outcome of football matches and other sports, many different approaches regarding dependency have been investigated. We propose the use of copula regression via the powerful GJRM (Generalised Joint Regression Models) framework and present its use for modelling match results. Motivated by the application to football and FIFA World Cups, in particular, we introduce two types of useful penalties. The first tackles a very specific issue occurring in sports tournaments and leagues (or other competitive situations), while the second is a Lasso-approximation yielding general sparsity.

E1543: **Football strategies via the effective playing space and tracking data**
Presenter: **Dimitris Karlis**, RC Athens University of Economics and Business, Greece
Co-authors: Marius Oetting

Modern analysis in football makes use of tracking data, i.e. we know exactly the position of the players and the ball with a high frequency. Such data can help coaches and scouts in several aspects, including game strategy and tactics, player evaluation, goal analysis, judging referee decisions, and talent identification, to name a few. We consider tracking data and focus on a metric that measures the position of the team on the field aiming at examining several tactical aspects. The metric is the convex hull created by the players of a team excluding the goalkeeper. It is also referred to as Effective Playing Space (EPS), calculated as the surface area (in square meters) of the convex hull of all players (excluding goalkeepers). For our analysis, we consider a novel hidden Markov model (HMMs) for modelling the EPS time series data jointly for the two teams, as they naturally accommodate the idea of a match progressing through different phases, with potentially changing tactics. The unobserved states in our HMM serve for the underlying tactics of a team (e.g. defensive vs. offensive style of play). The model enables to deepen the insights into the tactics of the teams, as interactions between them can be modelled additionally.

EO641 Room K2.31 Nash (Hybrid 07) STATISTICAL CHALLENGES IN COVID-19 EPIDEMIOLOGY Chair: Shaun Seaman

E1003: **Calibrating a large-scale stochastic meta-population model of COVID-19 in England and Wales**
Presenter: **Trevelyan McKinley**, University of Exeter, United Kingdom

Calibration of complex stochastic infectious disease models is challenging. These often have high-dimensional input spaces, with the models exhibiting complex, non-linear dynamics. Coupled with this is a paucity of necessary data, resulting in a large number of hidden states. Methods based on simulating the hidden states directly from the model-of-interest have the advantage that they are often easier code than likelihood-based approaches, and thus models can be developed and adapted quickly. However, they can be extremely computationally intensive, often requiring very large numbers of simulations in order to adequately explore the input space; rendering them infeasible for many large-scale problems. We extend recent developments in emulation-based calibration methods to calibrate a large-scale, stochastic, age-structured meta-population model of COVID-19 transmission in England and Wales. This approach, called history matching, develops quick surrogate models (emulators) that can be used in place of the complex simulator to efficiently explore the input space, and remove parts of the space where model fits are unlikely to be found. It does this whilst accounting for important sources of uncertainty: such as observation error, simulator uncertainty, emulator uncertainty and model discrepancy. We discuss various challenges relating to the model in question, and discuss the feasibility of implementing these methods for future pandemic modelling efforts.

E1050: **Estimating length of stay in a multi-state model conditional on the pathway, with application to patients with Covid-19**
Presenter: **Ruth Keogh**, London School of Hygiene and Tropical Medicine, United Kingdom

Multi-state models are used to describe how individuals transition through different states over time. The distribution of the time spent in different states, referred to as 'length of stay', is often of interest. Methods for estimating expected length of stay in a given state are well established. The focus will be on estimating the distribution of the time spent in different states conditional on the complete pathway taken through the states, which we call 'conditional length of stay'. The motivation comes from questions about the length of stay in hospital wards and intensive care units among patients hospitalised due to Covid-19. Conditional length of stay estimates are useful as a way of summarising individuals' transitions through the multi-state model, and also as inputs to mathematical models used in planning hospital capacity requirements. We will outline describe non-parametric methods for estimating the conditional length of stay distributions in a multi-state model in the presence of censoring. The methods will be illustrated using data on 42980 individuals hospitalised due to Covid-19 in the UK from March to July 2020, from the COVID19 Clinical Information Network.

E0669: **Estimating the severity of SARS-CoV-2 in England**
Presenter: **Anne Presanis**, MRC Biostatistics Unit, University of Cambridge, United Kingdom

The severity of the SARS-CoV-2 epidemic can be described by a number of different population-level measures summarising individual-level COVID-19 disease (clinical) severity and use of healthcare services, including the case-hospitalisation risk, hospitalised case-fatality risk, length of stay in hospital and infection-fatality risk amongst others. Some of these quantities are not directly measurable, due to the inherent inability to observe infections, rather than diagnoses, and are not straightforward to estimate, due to biases in observational data, reporting delays and other challenges. I will give an overview and examples of these challenges of estimating severity using a variety of survival, multi-state modelling and evidence synthesis approaches, including: jointly estimating the hospitalised case-fatality risk and length of stay in hospital; estimating the relative risk of hospital admission in Delta compared to Alpha cases; and combining evidence at different levels of severity to estimate the whole severity process, from infection to symptomatic infection to hospital admission to ICU admission to mortality.

E0423: **Adjusting for time of infection or positive test when estimating the risk of a post-infection outcome in an epidemic**
Presenter: **Shaun Seaman**, University of Cambridge, United Kingdom
Co-authors: Tommy Nyberg, Christopher Overton, David Pascall, Anne Presanis, Daniela De Angelis

When comparing the risk of a post-infection binary outcome, e.g. hospitalisation, for two variants of an infectious pathogen, it is important to adjust for calendar time of infection to avoid the confounding that would occur if the relative incidence of the two variants and the variant-specific risks of the outcome both change over time. Infection time is typically unknown, and the time of positive test is used instead. Likewise, time of positive test may be used instead of infection time when assessing how the risk of the binary outcome changes over calendar time. We show that if the mean time from infection to positive test is correlated with the outcome, the risk conditional on positive test time depends on whether the incidence of infection is increasing or decreasing over calendar time. This complicates the interpretation of risk ratios adjusted for positive test time. We also propose a simple sensitivity analysis that indicates how these risk ratios may differ from the risk ratios adjusted for infection time.

EC850 Room K0.16 (Hybrid 02) CONTRIBUTIONS IN STATISTICAL MODELLING II Chair: Pavlo Mozharovskyi

E0387: **Longitudinal modeling of age-dependent latent traits**
Presenter: **Oystein Sorensen**, University of Oslo, Norway
Co-authors: Anders Fjell, Kristine Walhovd

Latent variables are indispensable when measured responses reflect some underlying latent traits of interest. Often, latent traits depend smoothly on, e.g., age and location, with functional shapes hard to specify a priori. However, most latent variable models require parametric forms for both latent and observed variables, and flexible semiparametric models have limitations on the number of grouping levels or rely on restrictive assumptions like discrete-time. We present generalized additive latent and mixed models (GALAMM), extending generalized linear latent and mixed models (GLLAMM) by allowing both observed and latent variables to depend smoothly on observed variables. GALAMMs retain the flexibility offered by GLLAMMs, including an arbitrary number of grouping levels and the ability to fit a large number of response types. We show that any model in the GALAMM framework can be represented as a nonlinear mixed model and estimated by maximum likelihood. We compare algorithms for model fitting, and derive expressions for asymptotic covariance matrices. The motivating applications came from cognitive neuroscience, in which both latent cognitive abilities and structural characteristics of the brain follow smooth nonlinear trajectories across the lifespan, and we present examples where GALAMMs enabled answering research questions more easily than currently used tools.

E1584: **Investigating a time-varying degrees of freedom parameter in the robust modelling of longitudinal data**
Presenter: **Melanie Campbell**, Queen's University Belfast, United Kingdom

Longitudinal data is commonly found in a medical setting, for example, monitoring patients following the progression of disease or response to treatment. Linear mixed-effects models are one of the most popular models for representing longitudinal data. The model contains both fixed and random effects allowing for the within-subject correlation to be accounted for. Regrettably, the standard model is sensitive to outliers due to the fact that both the random errors and random effects are assumed to be normally distributed. The relatively untouched area of robust mixed modelling attempts to deal with this problem to minimise the risk of biased results when outliers are not accounted for. The robust models replace the typical Gaussian assumptions with t-distributional assumptions. Most work to date has both the random error and random effects modelled with a single degrees-of-freedom parameter. Yet it has been recognised that following treatment, for example, or other contributing factors, an outlying individual or the trend of an individual's measurement may conform to the population trends over time or conversely become outlying. This is the motivation behind a time-varying approach. Recent work has utilised splines to model the variation of outliers over time. The purpose is to expand more on these concepts and investigate the need for time-varying degrees of freedom parameter within a robust mixed model context.

E1638: **Doubly online changepoint detection for monitoring health status during sport activities**
Presenter: **Mattia Stival**, University of Padova-Dipartimento di Scienze Statistiche, Italy
Co-authors: Mauro Bernardi, Petros Dellaportas

An online framework is provided to analyze data recorded from smartwatches during running activities. In particular, we focus on identifying variations in the behavior of one or more measurements caused by physical condition changes such as physical discomfort, periods of prolonged de-training, or even malfunction of measuring devices. The framework considers data as a sequence of running activities, where one activity is a time series collecting over time physical and biometric data. We combine classical changepoint detection models with an unknown number of components to Gaussian state-space models to detect distributional changes between a sequence of activities (multivariate time series). The model considers multiple sources of dependence due to the sequential nature of subsequent activities, the autocorrelation structure within each activity, and contemporaneous dependence between different variables. We provide an online Expectation-Maximization (EM) algorithm involving a sequential Monte Carlo approximation of changepoint predicted probabilities. As a byproduct of our model assumptions, the proposed approach processes sequences of multivariate time series in a doubly online framework. While classical changepoint models detect changes between subsequent activities, the state space framework coupled with the online EM algorithm provides the additional benefit of estimating real-time probabilities that a single activity is a changepoint.

E1224: **An unrestricted MIDAS Poisson regression model**
Presenter: **Talha Omer**, Jonkoping University, Sweden, Sweden
Co-authors: Par Sjolander, Kristofer Mansson, BM Golam Kibria

An unrestricted mixed data sampling (U-MIDAS) model estimated using maximum likelihood (ML) for the Poisson regression model is proposed. An issue when using the standard U-MIDAS model is the overfitting due to a potentially large number of lags of the high-frequency variable used to predict the low-frequency regressand. Therefore, as a remedy, we suggest a regularized ridge approach. In terms of mean square error (MSE), we analytically prove the superiority of the ridge approach over the ML approach. Moreover, in a simulation study, we demonstrate that the ridge approach is superior in finite samples.

EC855 Room Virtual R30 CONTRIBUTIONS IN HIGH-DIMENSIONAL DATA ANALYSIS Chair: Mauro Bernardi

E1633: **Fast Bayesian model selection algorithms for linear regression models**
Presenter: **Mauro Bernardi**, University of Padova, Italy
Co-authors: Manuela Cattelan, Claudio Busatto

The issue of model selection for high-dimensional linear regression has been primarily addressed by assuming hierarchical mixtures as prior distributions. A spike component with Dirac probability mass at zero is introduced to exclude irrelevant covariates, thereby leading to Bayesian selection procedures that rely on the computation of the marginal posterior distribution for alternative model configurations. The exploration of the space of competing models is usually performed by means of computationally intensive simulation-based techniques. We address the issue of fast updating the variance-covariance matrix of the posterior distribution and the marginal posterior density itself, after a modification of the current design matrix. First, leveraging a thin QR factorization, novel algorithms to update the posterior variance-covariance matrix are proposed which avoid storage and update the Q matrix thus allowing noticeable savings. Then, the issue of evaluating the marginal posterior is considered, as it represents the bottleneck of any Bayesian model selection procedure. It is shown that the computation of the marginal posterior relies on the inverse of the R matrix, hence we develop a new methodology to update both this inverse and the related marginal posterior after the modification of the current design matrix. These methods do not need computationally intensive inversions of large dimensional matrices when performing marginal posterior evaluations.

E0348: **Learn2Evaluate: Predictive performance estimation with learning curves**
Presenter: **Jeroen Goedhart**, Amsterdam UMC, Netherlands
Co-authors: Mark van de Wiel

In high-dimensional prediction settings, i.e. when $p > n$, it remains challenging to estimate the test performance (e.g. AUC). Conventional resampling methods aim to balance between enough samples to reliably learn the model and estimate its performance. We show that combining estimates from a trajectory of subsample sizes, rendering a learning curve, leads to several benefits. Firstly, the use of a smoothed curve can improve the performance point estimate. Secondly, a still-growing- or saturating learning curve indicates whether or not additional samples will boost the prediction accuracy. Thirdly, comparing the trajectories of different learners results in a more complete picture than doing so at one sample size only. Fourthly, the learning curve allows computation of a useful lower confidence bound for the predictive performance. Standard cross-validation suffers from a limited amount of test samples, whereas the learning curve finds a better trade-off between training- and test sample sizes. This confidence bound is proven to be valid. We show coverage results from a simulation, and compare those to a state-of-the-art technique based on asymptotics and bootstrapping. Finally, we demonstrate the benefits of our approach by applying it to several classifiers of tumor location from blood platelet RNAseq data.

E0757: **Co-data learning in ridge models for high-dimensional data**
Presenter: **Mirrelijn van Nee**, Amsterdam University Medical Centers, Netherlands
Co-authors: Mark van de Wiel

Prediction is hard when data are high-dimensional, but additional information, like domain knowledge and previously published studies, may be helpful to improve predictions. Such complementary data, or co-data, provide information on the covariates, such as genomic location or p-values from external studies in cancer genomics. We use multiple and various co-data to define possibly overlapping or hierarchically structured groups of covariates. These are then used to estimate adaptive multi-group ridge penalties for generalised linear and Cox models. Available group adaptive methods primarily target settings with few groups, and therefore likely overfit for non-informative, correlated or many groups, and do not account for known structure on group level. To handle these issues, our method combines empirical Bayes estimation of the hyperparameters with an extra level of flexible shrinkage. This renders a uniquely flexible framework as any type of shrinkage can be used on the group level. We describe various types of co-data and propose suitable forms of hypershrinkage. The method is very versatile, as it allows for integration and weighting of multiple co-data sets, the inclusion of unpenalised covariates and posterior variable selection. As an illustrating example, we demonstrate the method in an oncogenomics setting.

E1251: **Simultaneous shrinkage of the linear mixed model's fixed and random effects using empirical Bayes**
Presenter: **Matteo Amestoy**, Amsterdam University Medical Centers, Netherlands
Co-authors: Mark van de Wiel, Wessel van Wieringen

Estimation of linear mixed models (LMM) from high dimensional data or studies with an imbalanced design requires regularization to prevent overfitting. However, standard LMM solvers do not facilitate shrinkage, while Bayesian hierarchical model implementations require fully parametric specification of the priors. The choice of the distributional form of these priors is guided by mathematical convenience. But an informed choice of the prior's parameter, especially for the covariance matrix of the random effects, is usually not at hand. We propose to select the hyperparameters in a data-driven fashion. Hereto we present an empirical Bayes (EB) method for the joint estimation of the parameters of a prior on the fixed effects and on the covariance matrix of the random effects. In our EB procedure, we maximize the marginal likelihood of the model using a Laplace approximation, where the model's maximum a posteriori is estimated with an expectation-maximization algorithm. We extensively compare the performance of our proposed method to standard LMM algorithms in simulation. Various scenarios show that our method improves the accuracy of the estimates and increases the prediction power of the LMM. Overall, estimation of the LMM from high-dimensional data benefits from the use of EB methods for data-driven regularization.

EG063 Room Virtual R29 CONTRIBUTIONS IN STATISTICAL METHODS FOR APPLICATIONS Chair: Daniel Gaigall

E0228: **Predicting inner temperature, humidity, and weight of beehives by using VAR models and sensorization**
Presenter: **Maria del Carmen Robustillo Carmona**, University of Extremadura, Spain
Co-authors: Carlos Javier Perez Sanchez, M Isabel Parra Arevalo

Bees play an important role in both agriculture and the environment. Precision beekeeping is a useful tool for predicting the state of a hive, which can anticipate different events such as hive collapse, swarming or the disease presence. The objective is to predict the hive's weight, temperature, and humidity using sensor data and meteorological information. For this purpose, data obtained by the we4bee project in the hive of Grund-und Mitteschule Vohburg have been used. The studied dataset collects information about internal temperature and humidity, weight, external weather conditions and some information provided by the beekeeper. VAR models were considered to predict the internal state of the beehive (temperature, humidity, and weight) employing historical data series that include exogenous meteorological data. Predictions were made for one, three and seven days ahead. To validate this model, a 100-fold cross-validation was performed, obtaining a mean absolute error (mean +- standard deviation) of 0.156 +- 0.137 kg in weight predictions, 0.987 +- 0.663 C in temperature and 2.925 +- 2.437 % in humidity. Given the promising results, we establish a starting point for predicting the state of the hives, which has not been addressed enough up to now.

E1513: **Analysing recreational boating traffic data in a panel time series data modelling setting**
Presenter: **Ebenezer Afrifa-Yamoah**, Edith Cowan University, Australia
Co-authors: Stephen M Taylor, Ute Mueller

The lack of continuity in recreational fisheries data due to intermittent sampling or surveys make trend estimation of effort difficult. Using a pooled time-series cross-sectional study design, panel generalized linear modelling techniques were used to identify potential determinants of recreational boating effort. Panels of data comprised time-lapse camera monitoring data of recreational boating effort and climatic and calendar-based variables over the period of 2011-2016 for four locations in different bioregions of Western Australia. Long-term equilibrium effects of the predictors on recreational boating traffic were estimated for within, between, random and pooled effects Poisson models. Significant effects ($p < 0.001$) on recreational boating traffic were observed for temperature, wind speed and direction, precipitation, sea level pressure, humidity, time of day and

month of the year across all panels. Sub-panel analyses revealed varying levels of importance of predictors with respect to locations. Non-linear tests of causality using artificial neural networks (ANN) based on vector auto-regressive neural network established significant unidirectional Granger causality of the study predictors on the recreational boating traffic.

E0717: **A sparse estimation method for sensory evaluation data with taking individual scaling differences into account**
Presenter: **Hironori Satomura**, Osaka University, Japan

In the sensory evaluation field, response styles, especially individual scaling differences, are frequently of concern. The traditional ways of analyzing sensory data, such as two-way ANOVA model comprising test stimuli as fixed effect and assessor related terms as random effects, are not capable of handling this scaling heterogeneity when investigating the differences among the test stimuli, which is of sensory scientists' interest. Assessor model, in which a multiplicative term introduced as an extra term in the two-way ANOVA model is, therefore, often utilized in the field. However, there still exists a limitation that the difference between each stimulus is examined via posthoc multiple comparisons, which does not necessarily produce a non-overlapping grouping of the stimuli. In order to tackle this problem, we propose a penalized likelihood method for this assessor model that encourages exact clustering of stimulus by taking the scaling difference into account. The model parameters are estimated through the EM algorithm with alternating direction method of multipliers. The usefulness of the proposed method is demonstrated through numerical examples.

E1597: **Liu after random forest: Application of machine learning methods in modeling high-dimensional chemical data**
Presenter: **Mohammad Arashi**, Ferdowsi University of Mashhad, Iran
Co-authors: Adewale F Lukman, Zakariya Y Algamal

In the modern era, using advanced technology, we have access to data with many features and therefore feature engineering has become a vital task in data analysis. One of the challenges in model estimation is to combat multicollinearity in high-dimensional data problems where the number of features exceeds the number of samples. We propose a novel, yet simple, strategy to estimate the regression parameters in a high-dimensional regime in the presence of multicollinearity. The proposed approach enjoys the good properties of the random forest and the simple structure of a class of linear unified estimators. We give a fast and straightforward algorithm to estimate the regression coefficients when multicollinearity exists. Numerical investigation reveals the superior performance of the method in prediction error. The technique is also applied to melting chemical data, where we conducted an estimation among 4885 features and discussed advantages.

CI012 Room K E. Safra (Multi-use 01) RECENT ADVANCES IN HIGH-DIMENSIONAL STATISTICS (VIRTUAL) Chair: Mark Podolskij

C0162: **Central limit theorems in high-dimensions: Recent developments**
Presenter: **Yuta Koike**, University of Tokyo, Japan

The purpose is to review recent progress in multivariate normal approximation on hyper-rectangles in the high-dimensional setting, where the dimension can be much larger than the sample size. Such an approximation is useful for justifying bootstrap approximations of maximum statistics in high-dimensional settings. It is, therefore, important for uniform inference in high-dimensional models.

C0163: **Dirft estimation for high dimensional diffusion models**
Presenter: **Mark Podolskij**, University of Luxembourg, Luxembourg

High dimensional estimation problems for continuous diffusion models are investigated. Parametric and non-parametric methods for drift and volatility estimation in fixed dimensions are nowadays well understood in the statistical and econometric literature. However, many diffusion systems appearing in practical applications are high dimensional, and thus a new estimation approach is required. We consider a LASSO type estimator of the drift function when complete paths observations are given. We derive rates of convergence for the estimator under sparsity constraints on the parameter.

C0161: **Quantitative limit theorems and bootstrap approximations for empirical projection**
Presenter: **Moritz Jirak**, University of Vienna, Austria

Given a sequence of random variables in some Hilbert space, the problem of finding distributional approximations for empirical projections is considered based on the empirical covariance operator and its spectral decomposition. Previously, this problem has been studied to a large degree only in a Gaussian setup. Novel quantitative limit theorems and bootstrap approximations are presented subject only to rather mild moment conditions. In many cases, these results approve upon even their Gaussian counterparts.

CO052 Room K0.50 (Hybrid 06) APPLIED FINANCIAL ECONOMETRICS Chair: Emese Lazar

C0358: **Testing stability in event observations with applications to IPO Performance**
Presenter: **Shixuan Wang**, University of Reading, United Kingdom
Co-authors: Lajos Horvath, Zhenya Liu, Gregory Rice, Yaosong Zhan

Many sequentially observed functional data objects are observable only at the times of certain events. For example, the trajectory of stock prices of companies after their initial public offering (IPO) can be observed when the offering occurs, and the resulting data may be affected by changing circumstances. It is of interest to investigate whether the mean behaviour of such functions is stable over time, and if not to estimate the times at which apparent changes occur. Since the frequency of events fluctuates each day, we propose a change point analysis that is comprised of two steps. In the first step, we segment the series into segments in which the frequency of events is approximately homogeneous using a new binary segmentation procedure for event frequencies. After adjusting the observed curves in each segment based on the frequency of events, we proceed in the second step by developing a method to test for and estimate change points in the mean of the observed functional data objects. We establish the consistency and asymptotic distribution of the change point detector and estimator in both steps, and study their performance using Monte Carlo simulations. An application to IPO performance data illustrates the proposed methods.

C1636: **Forward-looking market risk premium and its economic implications**
Presenter: **Shuyuan Qi**, University of Reading, United Kingdom
Co-authors: Emese Lazar, Radu Tunaru

The forward-looking market risk premium (FMRP) is a function of investors risk aversion and forward-looking volatility, skewness, and kurtosis of cumulative return. Using the S&P 500 index returns and VIX, we estimate the monthly FMRP from 1999 to 2020. We find that the FMRP estimated from the stochastic volatility model with a mean-reversion variance process adequately reflects market conditions and is always positive. We also find that the FMRP is significantly positively linked to the future market sentiment and significantly negatively linked to the future growth of real economic activities. Moreover, the market excess returns increase with FMRP under good market conditions and decrease with FMRP under bad market conditions.

C1715: **Forecasting VVIX using forecast combinations and LASSO**
Presenter: **Yushuang Jiang**, Peking University, United Kingdom
Co-authors: Emese Lazar

Motivated by the success of forecast combinations and the LASSO-type shrinkage methods, we attempt to answer the following question: is there an optimal VVIX forecasting method? If yes, then is this based on forecast combinations or LASSO? We show that forecast combinations perform best. We compare the forecasting performance of three individual models, eight combining methods and two LASSO-type models out-of-sample.

The results show that the simple median combining method delivers the lowest forecasting errors across the years. In addition, we discuss the model selection results of two shrinkage methods. Interestingly, instead of daily changes in the VVIX, the changes in monthly VVIX are key to predicting the VVIX.

C1696: **A discrete-time hedging framework with multiple factors and fat tails: On what matters**
Presenter: **Alex Badescu**, University of Calgary, Canada
Co-authors: Maciej Augustyniak, Jean-Francois Begin

A quadratic hedging framework is presented for a general class of discrete-time affine multi-factor models and investigates the extent to which multi-component volatility factors, fat tails, and a non-monotonic pricing kernel can improve the hedging performance. A semi-explicit hedging formula is derived for our general framework which applies to a myriad of the option pricing models proposed in the discrete-time literature. We conduct an extensive empirical study of the impact of modelling features on the hedging effectiveness of S&P 500 options. Overall, we find that fat tails can be credited for half of the hedging improvement observed, while a second volatility factor and a non-monotonic pricing kernel each contribute to a quarter of this improvement. Moreover, the study indicates that the added value of these features for hedging is different than for pricing. A robustness analysis shows that a similar conclusion can be reached when considering the Dow Jones Industrial Average. Finally, the use of a hedging-based loss function in the estimation process is investigated in an additional robustness test, and this choice has a rather marginal impact on hedging performance.

CO032 Room Virtual R28 HIGH DIMENSIONALITY, REGIME SHIFTS AND ROBUST INFERENCE Chair: Artem Prokhorov

C1217: **Using BIC-based forward stepwise instead of Lasso for Neyman-orthogonal estimation**
Presenter: **David M Drukker**, Sam Houston State University, United States
Co-authors: Di Liu

High-dimensional models that include many covariates which might potentially affect an outcome are increasingly common. A lasso-based approach and a stepwise-based approach to valid inference for a high-dimensional model are reviewed. Several important extensions are then discussed that make the estimators more usable in practice. Finally, Monte Carlo evidence is presented to help applied researchers choose which of several available estimators should be used in practice. The Monte Carlo evidence shows that our extensions to the literature perform well. It also shows that a BIC-stepwise approach performs well for a data-generating process for which the lasso-based approaches and a testing-stepwise approach fail. The Monte Carlo evidence also indicates the BIC-based lasso and plugin-based lasso can produce better inferential results than the ubiquitous CV-based lasso. Easy-to-use Stata commands are available for all the methods that we discuss.

C1298: **On the expectation and bias of the Gini coefficient under grouping**
Presenter: **Victor de la Pena**, Columbia University, United States

An approach is presented to calculate the Gini Coefficient of Income Inequality. We will provide an example using the Gamma Distribution. The bias incurred by grouping will be discussed as well.

C1443: **Change point detection in time series using mixed integer programming**
Presenter: **Alexander Semenov**, University of Florida and Saint Petersburg State University, United States
Co-authors: Anton Skrobotov, Artem Prokhorov

Recent advances in mixed-integer optimization (MIO) methods are used to develop a framework for the identification and estimation of structural breaks in time series. The framework requires a transformation of the classical structural break detection problem into a Mixed-Integer Quadratic Programming problem. MIO is capable of finding provably optimal solutions to this problem using a well-known optimization solver. The framework allows determining the unknown number of structural breaks. In addition to that, we demonstrate how to accommodate a specific required number of structural breaks, or a minimum required number of breaks. We demonstrate the effectiveness of our approach through extensive numerical experiments on synthetic and real-world data. We examine optimal and sub-optimal solutions to the problem, and the effect of tuning the parameters. We show how to choose the tuning parameters and compare our results with established econometric methods.

C1276: **On the compatibility of experts' Lorenz curves with plausible assumptions on unit nonresponse in household surveys**
Presenter: **Rami Tabri**, The University of Sydney, Australia
Co-authors: Sarah Dahmann, Brendan Beare

Many datasets from household surveys used in studies on income inequality of a reference household population suffer from unit nonresponse. A widely shared view on such surveys is that the missing incomes and households tend to belong to the top of the income distribution, which contrasts with the standard practice by survey designers who implement so-called corrections that assume unit nonresponse as missing completely at random. Consequently, a variety of approaches that adjust observed income distributions have been proposed by distributional experts to mitigate the bias in inequality estimates from such surveys. However, as with this practice by survey designers, such expert adjustments may be incompatible with plausible assumptions on unit nonresponse, which raises the question of how to justify using a particular expert's adjustment method in practice. A statistical procedure is proposed that helps practitioners answer this question. We derive bounds on the population Lorenz curve under different plausible unit nonresponse mechanisms and propose a design-based inference procedure for assessing the compatibility of an expert's adjustment of the observed Lorenz curve with those mechanisms. Finally, these bounds and the inference procedure are illustrated using the Household, Income and Labour Dynamics in Australia Survey.

CO882 Room Virtual R31 ENERGY ECONOMETRICS Chair: Malvina Marchese

C0283: **A new stochastic model for electricity prices**
Presenter: **Angelica Gianfreda**, University of Modena and Reggio Emilia, Italy
Co-authors: Derek Bunn

The wide range of models needed to support the various short-term operations for electricity generation demonstrates the importance of accurate specifications for the uncertainty in market prices. This is becoming increasingly challenging, since hourly price densities for electricity exhibit a variety of shapes, with their characteristic features changing substantially within the day and evolving over time. Furthermore, the influx of renewable power, wind and solar in particular, has made these density shapes very weather dependent. We develop a general four-parameter stochastic model for hourly prices, in which the four moments of the density function are dynamically estimated as latent state variables and furthermore modelled as functions of several plausible exogenous drivers. This provides a transparent and credible model that is sufficiently flexible to capture the shape-shifting effects, particularly with respect to the wind and solar output variations causing dynamic switches in the upside and downside risks. Extensive testing on German wholesale price data, benchmarked against quantile regression and other models in out-of-sample backtesting, validated the approach and its analytical appeal.

C0682: **Crude oil hedging using a regime-switching model**
Presenter: **Ioannis Moutzouris**, City, University of London, United Kingdom
Co-authors: Anastasios Zalachoris, Mahmoud Fatouh, Nikolaos Papapostolou, Panos Pouliasis

A regime-switching model is proposed for determining the optimal hedge ratio in the Brent crude oil market. The suggested framework extends the constant optimal hedge ratio model, employing four different crude oil market states, defined according to certain financial and economic metrics.

To measure its performance, the model is evaluated against four standard strategies (the nave 1:1 approach; the constant optimal hedge ratio; a time-varying optimal hedge ratio; and a 2-state Markov regime-switching model), in terms of variance reduction against the unhedged position benchmark. Our findings suggest that the proposed model outperforms the remaining hedging techniques by 36-51%, while it reduces the unhedged position's variance by 95%. Finally, the economic and policy-making implications of those findings are discussed.

C0580: **Modelling oil price risk using option- and forward market information**
Presenter: **Morten Risstad**, NTNU, Norway
Co-authors: Marie-Helene Gagnon, Gabriel Power, Sjur Westgaard

The dynamics of the historical return distribution for oil is complex with changing volatility, skewness (from negative to positive), and kurtosis over time. This is mainly due to changing markets expectations of future supply and demand conditions. We model the return distribution of oil prices as a function of implied volatility, skewness, kurtosis, and the shape of the oil forward curve. Quantile regression is applied with these measures as independent variables and oil price returns as the dependent variable. All measures are important in explaining the oil return distribution. Implied skewness and kurtosis have a significant effect explaining the tails of the distribution. The model performs very well when back-tested and compared to conventional risk models. Hence, the approach provides an excellent framework for understanding how risk drivers from the option and future market influence the oil price distribution.

C1466: **A mixed-frequency combination approach to forecast crude oil and gold future returns volatility and correlation**
Presenter: **Francesca Di Iorio**, University of Naples Federico II, Italy
Co-authors: Malvina Marchese, Ioannis Kyriakou, michael tamvakis

To forecast the covariance matrix of crude oil and gold futures returns, a novel forecast combination approach is proposed based on mixed information, i.e. high and low-frequency data. Specifically, the combination strategy identifies the optimal predictor using several loss functions via an iterative procedure based on the Model Confidence Set. The findings suggest that combined forecasts of multivariate GARCH and Realize Covariance models outperform each individual model and their equally weighted mean from a statistical as well as an economic perspective, indicating that low-frequency data improve volatility forecasting even when high-frequency data is available.

CO388 Room Virtual R32 TOPICS IN MODELING TIME SERIES AND PANEL DATA Chair: Markus Fritsch

C1170: **Discrepancy-based inference for intractable generative models using quasi-Monte Carlo**
Presenter: **Johanna Meier**, Leibniz University Hannover, Germany
Co-authors: Ziang Niu, Francois-Xavier Briol

Intractable generative models are models for which the likelihood is unavailable but sampling is possible. Most approaches to parameter inference in this setting require the computation of some discrepancy between the data and the generative model. This is, for example, the case for minimum distance estimation and approximate Bayesian computation. These approaches require sampling a high number of realisations from the model for different parameter values, which can be a significant challenge when simulating is an expensive operation. We propose to enhance this approach by enforcing "sample diversity" in simulations of our models. This will be implemented through the use of quasi-Monte Carlo (QMC) point sets. The key results are sample complexity bounds which demonstrate that, under smoothness conditions on the generator, QMC can significantly reduce the number of samples required to obtain a given level of accuracy when using three of the most common discrepancies: the maximum mean discrepancy, the Wasserstein distance, and the Sinkhorn divergence. This is complemented by a simulation study which highlights that improved accuracy is sometimes also possible in some settings which are not covered by the theory.

C1353: **Change-point analysis in high-dimensional econometric dynamic factor models**
Presenter: **Ansgar Steland**, RWTH Aachen, Germany

A new approach to test for a change in the covariance structure is proposed especially targeting high-dimensional econometric factor models. The approach is based on bilinear or quadratic forms of CUSUM statistics combined with a multiple testing procedure. Contrary to existing tests that fail and/or are computationally infeasible when it comes to high dimensions p, the proposed methodology can even be used for $p > T$. The class of factor models allowed for covers many specifications used in econometric data analysis and modeling. It even allows for an infinite number of correlated factors. The approach is carefully examined by simulations. We illustrate the approach by analyzing the impact of the Covid crash on the Fama-French factors.

C1546: **Fixed event forecasting of multiple lead times**
Presenter: **Markus Fritsch**, University of Passau, Germany
Co-authors: Harry Haupt, Joachim Schnurbus

Fixed event forecasts target a particular event at a specific future date. Examples are the outcomes of sports events, wind speeds of hurricanes when they hit inhabited areas, or the results of political elections. As rolling event forecasting, where each forecast refers to a different time period, is by far more common, we first detail the particularities of fixed event forecasting. Second, we review testable and theoretical conditions for the efficiency of fixed event forecasts. Third, a nonparametric framework for fixed event forecasting is developed and forecasting the results of German state elections is illustrated for multiple election cycles in selected states. We compare the results to traditional election polls and methods from modern time series analysis. Additionally, the benefits of averaging across multiple estimation windows are investigated for the different approaches. Our results suggest that off-the-shelf statistical models outperform conventional election polls, even on days when new polls become available.

C1571: **Properties of an IV-estimator based on aggregated nonlinear moment conditions**
Presenter: **Joachim Schnurbus**, University of Passau, Germany
Co-authors: Andrew Adrian Yu Pua, Markus Fritsch

An instrumental variables (IV) estimator based on aggregated nonlinear moment conditions is proposed in order to estimate the autoregressive parameter in linear dynamic panel data models. As the IV-estimator may converge to two distinct solutions, we provide a weighting scheme to identify the correct solution. The derivation of the large sample properties of the proposed estimator is underlined by Monte Carlo results.

CO046 Room Virtual R34 EMPIRICAL MODELS FOR CORPORATE FINANCE AND BANKING Chair: Leone Leonida

C1411: **Is there a underlying shape of the investment cash flow sensitivity**
Presenter: **Leone Leonida**, King's College London, United Kingdom

The shape of investment cash flow sensitivity (ICFS) with respect to the degree of financing constraints is a controversial issue in corporate finance research. We contribute to resolving this controversy by proposing an analysis of the shape that sidesteps all the uncertainties regarding the scheme used to sort firms according to financing constraints. The evidence suggests that the ICFS is non-monotonic, and the underlying ICFS is inverse basin shaped, regardless of the financing constraints metric. We show that the inverse basin shape encompasses all other shapes documented by previous studies, and it is engendered only by a non-monotonic relation between ICFS and the degree of financing constraints.

C1417: **Political replacement effect and financial development: Evidence across countries**
Presenter: **Alfonsina Iona**, Queen Mary University of London, United Kingdom

The Politics and Finance literature argues that political factors are responsible for shaping a country financial development. In line with this

view, we analyze the impact of political competition on financial development across 124 countries over the period 1970-2015. The results show that political competition Granger-causes financial development, and there is a political replacement effect of political competition on financial development. However, the relationship between political competition and financial development is U-shaped. The results are robust to the origin of a country legal system, to different sub-samples and alternative measures of financial development and political competition.

C1426: **Strategic financial networks**
Presenter: **Marina Dolfin**, King's College London, United Kingdom

The aim is to tie the observed dynamics of liquidity allocation of financial intermediaries to the inherent randomness of the structure of the underlying network, as well as to the strategic process of link formation driven by exogenous regulatory policies. The traditional focus of banking supervision is on individual institutions, but liquidity allocation shows the typical features of network-based phenomena in a complex economic environment. In normal times, the highly interconnected system of banks enhances liquidity allocation through interconnectedness and increase risk-sharing whilst, in times of crisis, interconnections lead to amplification of shocks. This property has been called a knife-edge, or robust-yet-fragile. It is worth analyzing which characteristics of complex systems correlate with a high degree of robustness and resilience. Even if there is a wide literature on algorithms designed to generate networks with desired properties, their rules are based on pure chance. This represents a strong limitation when modelling social and economic networks, because of missing incentives to create or delete relationships among actors on the network. The focus is on the dynamical balance between the overall societal welfare and individual incentives, then trying to include incentives in the process of link formation in order to analyse the global network properties that arise, by focussing on robustness and resilience.

C1427: **Market driven securitization**
Presenter: **Eleonora Muzzupappa**, Kingś College London, United Kingdom

How, and how much, does the performance of the stock market affect banks securitization activity? The analysis of a panel of EU and US banks shows that the former shapes the latter both directly and by interacting with some balance-sheet items. We find that the impact of the stock market performance upon the banks securitization, the channels with which it interacts with the balance sheet items and the sign that these impacts take depend upon the market discipline, that shapes both the banks business model of securitization, and the condition of the financial market.

CO667 Room Virtual R35 FINANCIAL ECONOMETRICS IN A BAYESIAN FRAMEWORK Chair: Yong Song

C0245: **Does the choice of realized covariance measures empirically matter? A Bayesian density prediction approach**
Presenter: **Jia Liu**, Saint Mary's University, Canada
Co-authors: Qiao Yang, Jin Xin

A new approach is suggested to evaluate realized covariance (RCOV) estimators via their predictive power on return density. By jointly modeling returns and RCOV measures under a Bayesian framework, the predictive density of returns and ex-post covariance measures are bridged. The forecast performance of a covariance estimator can be assessed according to its improvement in return density forecasting. Empirical applications to equity data show that several RCOV estimators consistently perform better than others and emphasize the importance of RCOV selection in covariance modeling and forecasting.

C0381: **Identification and forecasting of bull and bear markets using multivariate returns**
Presenter: **Yong Song**, University of Melbourne, Australia
Co-authors: John Maheu, Jia Liu

Bull and bear market identification generally focuses on a broad index of returns through a univariate analysis. A new approach is proposed to identify and forecast bull and bear markets through multivariate returns. The model assumes all assets are directed by a common discrete state variable from a hierarchical Markov switching model. The hierarchical specification allows the cross-section of state-specific means and variances to differ over bull and bear markets. We investigate several empirically realistic specifications that permit feasible estimation even with 100 assets. The results show that the multivariate framework provides competitive bull and bear regime identification and improves portfolio performance and density prediction compared to several benchmark models, including univariate Markov switching models.

C0389: **A Bayesian semiparametric stochastic volatility model with Markovian mixtures**
Presenter: **Qiao Yang**, ShanghaiTech University, China
Co-authors: John Maheu, Chenxing Li

A previous Bayesian semiparametric stochastic volatility (SV-DPM) model is extended. Instead of using a Dirichlet process mixture (DPM) for return innovations that capture a constant unknown density, we use an infinite hidden Markov model (IHMM). This allows for time variation in the return density beyond that attributed to latent volatility. The new model (SV-IHMM) also nests the SV-DPM as a special case and greatly improves the density forecast from the SV model with Student-t innovations (SVt) compared to the SV-DPM. The model is applied to several applications, and a comparison is made with the Dirichlet process version. The results show that SV-IHMM generally requires fewer states than the SV-DPM on average and predicts better out-of-sample. Furthermore, predictive densities from the SV-IHMM exhibit clear distributional shifts over time. These results are robust to different hyperparameter prior values.

C0497: **Product partitioned dynamic factor models: An application characterising Melbourne housing price co-movements**
Presenter: **Zhuo Li**, Monash Univeristy, Australia

Bayesian nonparametric methods are used to accommodate the random partition issue surrounding block dynamic factor models (DFMs). The DFM with a block structure assumes that time series within the same block co-move with cluster-specific factors, where series from distinct blocks only communicate through global factors. The premise of this modelling framework is that the cluster assignment is given *ex-ante*. In the presence of *partition uncertainty*, three novel modelling frameworks are proposed that feature a random partitioning process in dynamic factor models to endogenously determine the partition size and cluster memberships, where each proposed model is designed for its own purpose, since no single model suits all applications. From the simulation experiment, the in-sample fit results show that allowing for the random partitioning process over the block DFM provides additional precision in prediction. Such precision gain is further confirmed in the empirical study of Melbourne housing price co-movement analysis, for both in-sample and out-of-sample exercises. In addition, it is suggested that the effect of economic drivers on housing price dynamics may not be fully captured if suburbs-level partition uncertainty is ignored while accommodating spatial dependence.

CO108 Room Virtual R38 QUANTITATIVE INVESTMENT Chair: Serge Darolles

C0684: **Forecasting portfolio weights**
Presenter: **Hugues Langlois**, HEC Paris, France

A new methodology is proposed to implement unconditionally optimal dynamic mean-variance portfolios. We model portfolio allocations using an auto-regressive process in which the shock to the portfolio allocation is the gradient of the investors' realized certainty equivalent with respect to the allocation. The methodology can accommodate transaction costs, short-selling and leverage constraints, and a large number of assets. In out-of-sample tests using equity portfolios, long-short factors, government bonds, and commodities, we find that its risk-adjusted performance, net of transaction costs, is on average more than double that of other benchmark allocations.

C0750: **Market impact decay and capacity**
Presenter: **Hector Chan**, University Paris Dauphine, France

Recent studies have documented that market impact decays slowly through time. We study the impact of such slow decay on trading strategies' capacity. To do so, we propose a numerical methodology to estimate capacity. A key benefit of such a procedure is its flexibility in incorporating any specification of market impact. In particular, as traders tend to be more auto-correlated when capital devoted to a trading strategy increases, capacity is sensitive to assumptions on market impact decay. We show that incorporating market impact's slow decay leads to trading strategy capacity that is significantly lower than conveyed in previous capacity studies.

C0892: **Dissecting beta**
Presenter: **Costas Xiouros**, BI Norwegian Business School, Norway
Co-authors: Paul Ehling

In a framework where the CAPM holds conditionally, a model is developed with (fairly) general dividend dynamics and stochastic discount factor that accounts for standard asset pricing moments and the wide range of unconditional betas. The model features a strong time-varying cyclicality component in the dividend dynamics producing two effects: unconditional CAPM alphas are within statistical error and conditional betas are non-linear making it hard to estimate their relations to observable quantities. Consequently, stationary cash-flow dynamics such as the ones of industry returns may be generated in a mean-variance efficient fashion despite that industry characteristics help explain their betas.

C1390: **Forecasting option returns with news**
Presenter: **Gang Li**, The Chinese University of Hong Kong, Hong Kong
Co-authors: Bing Han, Jie Cao, Ruijing Yang, Xintong Zhan

The aim is to study whether text data contains useful information to forecast the cross-sectional equity option returns. We apply both lexicon-based and machine-learning approaches to extract quantitative signals from over six million news articles. The machine-learning methods outperform lexicon-based approaches in predicting the delta-hedged option returns, and generate sizable profits. The predictability is robust after controlling for volatility-related information and other known predictors. An analysis of the keywords identified by machine-learning methods suggests the predictability is largely related to sentiment. We highlight the importance of analyzing unstructured data like texts with machine learning approaches by examining the derivatives market.

CO470 Room K2.41 (Hybrid 09) DEEP CALIBRATION OF FINANCIAL MODELS Chair: Christa Cuchiero

C1103: **Robust pricing and hedging via neural SDEs**
Presenter: **David Siska**, University of Edinburgh, United Kingdom
Co-authors: Patryk Gierjatowicz, Marc Sabate-Vidales, Lukasz Szpruch, Zan Zuric

Modern data science techniques are opening the door to robust, data-driven model selection mechanisms. However, most machine learning models are "black-boxes" as individual parameters do not have a meaningful interpretation. In contrast, classical risk models based on stochastic differential equations (SDEs) with fixed parametrisation are well understood. Unfortunately, the risk of using an inadequate model is hard to detect and quantify. Instead of choosing a fixed parametrisation for the model SDE, we allow the drift and diffusion to be given by overparametrised neural networks. This allows one to find robust bounds for prices of derivatives and the corresponding hedging strategies. The resulting model, called neural SDE, is an instantiation of generative models and is closely linked with the theory of causal optimal transport. Neural SDEs allow consistent calibration under both the risk-neutral and the real-world measures. Thus the model can be used to simulate market scenarios needed for assessing risk profiles and hedging strategies. We develop and analyse novel algorithms needed for the efficient use of neural SDEs and we validate our approach with numerical experiments using both market and synthetic data. The code used is available at Github.

C1015: **Consistent recalibration models and deep calibration**
Presenter: **Matteo Gambara**, ETHZ, Switzerland
Co-authors: Josef Teichmann

Consistent Recalibration models (CRC) have been introduced to capture in necessary generality the dynamic features of term structures of derivatives' prices. Several approaches have been suggested to tackle this problem, but all of them, including CRC models, suffered from numerical intractabilities mainly due to the presence of complicated drift terms of consistency conditions. We overcome this problem by machine learning techniques, which allow storing the crucial drift term's information in neural network type functions. This yields first-time dynamic term structure models which can be efficiently simulated. As a side result, we are able to simulate for an indefinite time the evolution of an implied volatility surface under no-arbitrage constraints.

C1024: **Universal signature-based models**
Presenter: **Sara Svaluto-Ferro**, University of Vienna, Austria
Co-authors: Christa Cuchiero, Guido Gazzani

Universal classes of dynamic processes based on neural networks and signature methods have recently entered the area of stochastic modeling and Mathematical Finance. This has opened the door to robust and more data-driven model selection mechanisms, while first principles like no-arbitrage still apply. We focus on signature SDEs whose characteristics are linear functions of a primary underlying process, which can range from a (market-inferred) Brownian motion to a general multidimensional tractable stochastic process. The framework is universal in the sense that any classical model can be approximated arbitrarily well and that the model characteristics can be learned from all sources of available data by simple methods. Indeed, we derive formulas for the expected signature in terms of the expected signature of the primary underlying process. These formulas enter directly in the calibration procedure to option prices, while time-series data calibration just reduces to a simple regression.

C1112: **Data-driven market simulators and some simple applications of signature kernel methods in mathematical finance**
Presenter: **Blanka Horvath**, TU Munich, Germany

Techniques that address sequential data have been a central theme in machine learning research in the past years. More recently, such considerations have entered the field of finance-related ML applications in several areas where we face inherently path-dependent problems: from (deep) pricing and hedging (of path-dependent options) to generative modelling of synthetic market data, which we refer to as a market generation. We revisit Deep Hedging from the perspective of the role of the data streams used for training and highlight how this perspective motivates the use of highly accurate generative models for synthetic data generation. Stochastic processes are at their core random variables with values on path space. However, while the distance between two (finite-dimensional) distributions was historically well understood, the extension of this notion to the level of stochastic processes remained a challenge until recently. We discuss the effect of different choices of such metrics while revisiting some topics that are central from a regulatory (and model governance) perspective.

CC875 Room K2.40 (Hybrid 08) CONTRIBUTIONS IN PORTFOLIO ANALYSIS Chair: Aleksey Kolokolov

C0414: **Impact of trading rules in portfolio process with respect to market risk capital requirement**
Presenter: **David Nedela**, VSB - TU Ostrava, Czech Republic

A permissible investment approach combines a general portfolio model with other disciplines in the financial area to find a suitable portfolio strategy for investment. The aim is to examine the impact of several trading rules including technical analysis and stochastic dominance approaches in the portfolio creation process in US and UK markets during different time horizons, capturing different market conditions. The analysis is mainly focused on the needs of market risk capital requirement based on the Basel III approach. We consider two strategies for implementing trading rules

in the portfolio creation process. Strategy 1 is based on eliminating the whole market systemic risk, represented by the proportion of assets that meet certain trading rules with the alternative investment in a risk-free asset. The second strategy is focused on the use of pure assets that meet specific trading rules. From the results, using strategy 1 to find systemic risk during the crisis reduces the riskiness of the portfolio and capital requirement with a similar level of profitability, whilst strategy 2 generates only slightly higher profit. In a period with a growing economy, strategy 2 is more profitable, but the required positive effect on the level of a capital requirement is not achieved.

C0918: **Robust hedging of long-term investments under interest rate risk and inflation risk**

Presenter: **Lieske Coumans**, Tilburg University, Netherlands

Co-authors: Anne Balter, Frank de Jong

Investors often hedge their liabilities against nominal interest rate risk. However, inflation risk also plays an important role in real wealth outcomes, especially in the long run. If both risks follow a bivariate mean-reverting process, optimal allocations in nominal bond strategies typically turn out to be extreme, in particular when the bond maturities lay close to each other. We show that this makes the investment strategy sensitive to small changes in the mean-reversion parameters and the feedback parameter that takes into account the impact of the inflation rate level on the nominal interest rate drift. We perform a numerical analysis to demonstrate that small estimation errors of these parameters might have a large impact on terminal real wealth. A range of values of the feedback parameter is applied to compare the resulting investment strategies to related literature. We find that the optimal two bond strategies involve one medium-term bond and one very long term bond, but these strategies are very sensitive to parameter uncertainty. One bond strategies are more robust, but cannot hedge inflation risk, which results in a large loss in the Certainty Equivalent Wealth of a very risk-averse investor.

C1590: **Quantile maximizer in action**

Presenter: **Martin Hronec**, UTIA AV CR vvi, Czech Republic

The out-of-sample performance of portfolio selecting investors with τ-quantile preferences is studied. Investor's risk aversion is captured by τ, where more risk-averse investor maximize lower τ-quantile. Using a number of empirical and simulated datasets, we document differences in optimal portfolios across different levels of risk aversion. We also compare optimal quantile portfolios with equal-weighting and global minimum variance portfolios, documenting heterogeneity in portfolio compositions as well as in the out-of-sample performance.

C1540: **Bond portfolio optimization in turbulent times: A dynamic Nelson-Siegel approach with Wishart stochastic volatility**

Presenter: **Richard Schnorrenberger**, Kiel University, Germany

Modeling and forecasting the time-varying volatility of bond yields play a prominent role in many finance applications. However, amid periods of financial turmoil, managing interest rate risk on a daily basis is rather a challenging task due to extreme realizations and sudden changes in bond yields that can easily lead to implausible density forecasts. To reduce forecasting uncertainty and account for structural instability in volatile bond markets, the predictive performance of yield curve factor models with time-varying VAR parameters and Wishart stochastic volatility is investigated under a Bayesian MCMC scheme. A bond portfolio optimization and Value-at-Risk forecasting application to daily US Treasury yields also highlight the potential of modeling frameworks with factor Wishart stochastic volatility. The results clearly indicate that the proposed modeling features are economically motivated due to their outperformance in terms of portfolio allocation and risk management during turbulent times including the Great Recession and COVID-19 pandemic.

Saturday 18.12.2021 14:20 - 16:00 Parallel Session D – CFE-CMStatistics

EI016 Room K E. Safra (Multi-use 01) DESIGN AND ANALYSIS OF EXPERIMENTS (HYBRID) Chair: Kalliopi Mylona

E0175: **Neighbour balance and evenness of distribution of treatment replications in row-column designs**
Presenter: **Hans-Peter Piepho**, Universitaet Hohenheim, Germany

Row-column designs allow error control in field experiments by blocking in two dimensions. While this strategy can capture spatial heterogeneity aligned with blocks and account for effects due to the farming operations along rows and columns, it suffers from the occasional clustered occurrence of several replications of the same treatment. This property of classical row-column designs has hampered their more widespread use in practice. A further issue of practical importance is the degree of neighbor balance of a design, that is, the frequency of adjacencies of pairs of treatments. Two design strategies are proposed to simultaneously optimize the evenness of spatial distribution of treatment replication and neighbor balance. Three examples are given to illustrate the proposed methods and demonstrate that both approaches yield comparable and satisfactory results.

E0169: **Factor selection in screening experiments**
Presenter: **John Stufken**, University of North Carolina at Greensboro, United States
Co-authors: Rakhi Singh

Screening designs are used in design of experiments when, with limited resources, important factors are to be identified from a large pool of factors. Typically, a screening experiment will be followed by a second experiment to study the effect of the identified factors in more detail. As a result, the screening experiment should ideally screen out a large number of factors to make the follow-up experiment manageable, without screening out important factors. The Gauss-Dantzig Selector (GDS) is often the preferred analysis method for screening designs. While there is ample empirical evidence that fitting a main-effects model can lead to incorrect conclusions about the factors if there are interactions, including two-factor interactions in the model increases the number of model terms dramatically and challenges the GDS analysis. We discuss a new analysis method, called Gauss Dantzig Selector Aggregation over Random Models (GDS-ARM), which aggregates the effects from different iterations of the GDS analysis using different randomly selected interactions columns each time.

E0733: **Active learning: Intelligent subsampling**
Presenter: **Jesus Lopez-Fidalgo**, University of Navarra, Spain
Co-authors: Alvaro Cia-Mina

The Big Data sample size introduces statistical and computational challenges to extract useful information from data sets. The subsampling procedure is widely used to downsize the data volume and allows computing estimators in regression models. Usually, subsampling is performed defining a weight for each point and selecting a subset according to these weights. The subsample can be chosen at random (Passive Learning), but in order to obtain better estimators, the optimal experimental design theory can be used to search for an influential sub-sample (Active Learning). This has been developed in the literature for linear and logistic regression, obtaining algorithms based on D-optimality and A-optimality. To the authors' knowledge, the distribution of the explanatory variables has never been considered for obtaining a subsample. We study the effect of the explanatory variables distribution on the estimation as well as the optimal design. We first assume the normality of the covariates and later we measure the impact of skewness and kurtosis on the estimation and optimal designs. Then, we propose a novel method to obtain optimal subsampling through D-optimality, taking into account the marginal distribution of the covariates. The D-optimal design is computed by an exchange algorithm to obtain the subsample.

EO525 Room K0.16 (Hybrid 02) ASSOCIATION AND DEPENDENCE Chair: Johannes Wiesel

E1061: **Measuring association on topological spaces using kernels and geometric graphs**
Presenter: **Bodhisattva Sen**, Columbia University, United States

The aim is to propose and study a class of simple, nonparametric, yet interpretable measures of association between two random variables X and Y taking values in general topological spaces. These nonparametric measures – defined using the theory of reproducing kernel Hilbert spaces – capture the strength of dependence between X and Y and have the property that they are 0 if and only if the variables are independent and 1 if and only if one variable is a measurable function of the other. Further, these population measures can be consistently estimated using the general framework of graph functionals which include k-nearest neighbour graphs and minimum spanning trees. Moreover, a sub-class of these estimators are also shown to adapt to the intrinsic dimensionality of the underlying distribution. Some of these empirical measures can also be computed in near-linear time. Under the hypothesis of independence between X and Y, these empirical measures (properly normalized) have a standard normal limiting distribution. Thus, these measures can also be readily used to test the hypothesis of mutual independence between X and Y. In fact, as far as we are aware, these are the only procedures that possess all the above mentioned desirable properties.

E1026: **On boosting the power of Chatterjee's rank correlation**
Presenter: **Fang Han**, University of Washington, United States
Co-authors: Zhexiao Lin

Chatterjee's ingenious approach to estimating a previous measure of dependence based on simple rank statistics has quickly caught attention. This measure of dependence has the unusual property of being between 0 and 1, and being 0 or 1 if and only if the corresponding pair of random variables is independent or one is a measurable function of the other almost surely. However, more recent studies showed that independence tests based on Chatterjee's rank correlation are unfortunately rate-inefficient against various local alternatives and they call for variants. We answer this call by proposing revised Chatterjee's rank correlations that still consistently estimate the same dependence measure but provably achieve near-parametric efficiency in testing against Gaussian rotation alternatives. This is possible via incorporating many right nearest neighbors in constructing the correlation coefficients. We thus overcome the "only one disadvantage" of Chatterjee's rank correlation.

E0837: **How simplifying and flexible is the simplifying assumption in pair-copula constructions**
Presenter: **Sebastian Fuchs**, University of Salzburg, Austria
Co-authors: Thomas Mroz, Wolfgang Trutschnig

Motivated by the popularity and the seemingly broad applicability of pair-copula constructions underlined by numerous publications in the last decade, we tackle the unavoidable question of how flexible and simplifying the commonly used 'simplifying assumption' is from an analytic perspective and provide answers to two open questions. Aiming at the simplest possible setup for deriving the main results we first focus on the three-dimensional setting. We prove that the family of simplified copulas is flexible in the sense that it is dense in the set of all copulas with respect to the uniform metric. Considering stronger notions of convergence like the one induced by the metric D_1, by weak conditional convergence, by total variation, or by Kullback-Leibler divergence, however, the family even turns out to be nowhere dense and hence insufficient for any kind of flexible approximation. Furthermore, returning to the uniform metric we show that the partial vine copula is never the optimal simplified copula approximation of a given, non-simplified copula, and derive examples illustrating that the corresponding approximation error can be strikingly large. Moreover, the mapping assigning each copula its unique partial vine copula turns out to be discontinuous with respect to the uniform metric, implying a surprising sensitivity of partial vine copula approximations. The aforementioned main results are then extended to the general multivariate setting.

E0209: **Measuring association with Wasserstein distances**
Presenter: **Johannes Wiesel**, Columbia University, United Kingdom

Coupling between two probability measures on a Polish space is considered. We propose and study a class of nonparametric measures of association between these two measures. The analysis is based on the Wasserstein distance between the disintegration of the coupling with respect to the first coordinate and the marginals. We also establish basic statistical properties of this new class of measures: we develop a statistical theory for strongly consistent estimators and determine their convergence rate. Throughout our analysis, we make use of the so-called adapted/causal Wasserstein distance. Our class of measures offers an alternative to the correlation coefficient. In contrast to previous works, our approach also applies to probability laws in general Polish spaces.

EO236 Room K0.18 (Hybrid 03) TRACK: BAYESIAN SEMI- AND NON-PARAMETRIC METHODS II Chair: Bernardo Nipoti

E0683: **Bayesian graphical modelling for heterogeneous causal effects**
Presenter: **Federico Castelletti**, Universita Cattolica del Sacro Cuore (Milan), Italy
Co-authors: Guido Consonni

A Directed Acyclic Graph (DAG) provides an effective framework for analyzing causal relations among variables based on observational data. In particular, the effect on a response due to a hypothetical intervention on a variable in the system can be meaningfully addressed. We account for uncertainty on the DAG structure, and we overcome the usual assumption that the graph is common to all observations by allowing for heterogeneity of the underlying population. We consider a Dirichlet Process (DP) mixture of DAG models, where each component of the mixture is a Gaussian DAG model, with cluster-specific conditional independencies encoded by the DAG. Our methodology allows us to identify homogeneous subgroups of individuals, each having a distinct causal effect following an intervention on a target variable. The inference is based on Normal-DAG-Wishart priors for the mean and the Cholesky parameters of each Gaussian DAG model. Lack of identifiability of the underlying DAG due to observational data and computational issues are discussed. An application to protein expression data from Acute Myeloid Leukemia (AML) patients is presented, with the aim of producing estimates of cluster-specific causal effects on disease progression following interventions on targeted proteins.

E1080: **Bayesian mixture models for the prediction of extreme observations**
Presenter: **Isadora Antoniano-Villalobos**, Ca' Foscari University of Venice, Italy
Co-authors: Simone Padoan, Boris Beranger

In many applications with interest in large or extreme observations, usual inferential methods may fail to reproduce the tail behaviour of the variables involved. Recent literature has proposed the use of multivariate extreme value theory to predict an unobserved component of a random vector given large observed values of the rest. This is achieved through the estimation of the angular measure controlling the dependence structure in the tail of the distribution. The idea can be extended and used for the prediction of multiple components at adequately large levels, provided the model used for the angular measure is sufficiently flexible enough to capture complex dependence structures. The use of Bernstein polynomials ensures such flexibility and their interpretation as mixture models allows the use of current trans-dimensional MCMC posterior simulation methods for inference.

E1523: **Using random partition models for flexible change-point analysis in multivariate processes**
Presenter: **Garritt Page**, BYU, United States

Change point analyses are concerned with identifying positions of an ordered stochastic process that undergo abrupt local changes of some underlying distribution. When multiple processes are observed, it is often the case that information regarding the change point positions is shared across the individual processes. A method is described that takes advantage of this type of information. Since the number and position of change points can be described through a partition with contiguous clusters, the approach is based on developing a dependent model for these types of partitions. We describe computational strategies and illustrate improved performance in detecting change points through a small simulation study. We then apply the method to a financial data set of emerging markets in Latin America.

E0630: **Bayesian mixture models in regression with variable selection**
Presenter: **Aixin Tan**, University of Iowa, United States

Heterogeneous data are ubiquitous in scientific studies. In regression problems, the response in different subpopulations may be influenced by different subsets of covariates. We propose using mixtures, especially mixtures of finite mixtures (MFM), to model the joint distribution of the response and the covariates. In particular, we adopt a parameterization that explicitly involves vectors of regression coefficients within each subpopulation, each assigned spike-and-slab priors to achieve subpopulation-specific variable selection. MCMC algorithms are used for computing, leading to versatile posterior inferences, such as clustering, individual profiling, and predictions.

EO096 Room Virtual R18 RECENT ADVANCES IN MODEL-BASED CLUSTERING Chair: Pietro Coretto

E0704: **Nonparametric consistency for ML-based clustering with finite mixtures of elliptically symmetric distributions**
Presenter: **Pietro Coretto**, University of Salerno, Italy
Co-authors: Christian Hennig

The k-means method is often referred to as nonparametric, based on the nonparametric consistency theorem, which shows that without assuming any parametric model, under general conditions, the k-means solution converges to its own canonical functional (population version). We prove a similar result for a constrained MLE for finite mixtures of a general class of elliptically symmetric distributions (including popular models such as the Gaussian, Student-t, and Laplace, etc.). We also show that for data generated from distributions producing "clear clustering", the partition implied by the value of the ML canonical functional can be interpreted appropriately as corresponding to the clusters in the population.

E0857: **Group number identification in random projection ensemble model-based clustering**
Presenter: **Laura Anderlucci**, University of Bologna, Italy
Co-authors: Angela Montanari

A novel procedure is proposed for model-based clustering of high-dimensional data, based on random projection ensembles. Specifically, a Gaussian mixture model is fit to random projections of the high-dimensional data and a subset of solutions is selected accordingly to the Bayesian Information Criterion; the multiple 'base' results are then aggregated via consensus to obtain the final partition. The proposed algorithm is a very general tool for model-based clustering of high-dimensional data. We explore in detail its behaviour within the Gaussian mixture model framework only; however, many other distributions can in principle be used. The procedure is derived under the assumption that the number of clusters G is fixed and known. However, in real-life applications, it may happen that there is no insight about the 'true' number of homogeneous groups and such information has to be inferred from the data. Here, some model selection procedures are suggested as a valid tool to choose the number of clusters when such information is not available.

E1346: **Model-based co-clustering of multivariate time-dependent data**
Presenter: **Alessandro Casa**, University College Dublin, Ireland
Co-authors: Charles Bouveyron, Elena Erosheva, Giovanna Menardi

Multivariate time-dependent data arise when multiple features are measured for a set of units over different time instants. When dealing with such

data, flexible statistical tools are needed in order to account for characteristics such as the relations among both time observations and variables, the possible subject heterogeneity and arbitrarily shaped time evolutions. A new co-clustering strategy is outlined, grouping simultaneously variables and individuals, and being adequate both for longitudinal and functional data. The proposed approach relies on the shape invariant model which is embedded in the latent block model, representing the most popular model-based co-clustering strategy. To account for the specific features of the shape invariant model, the estimation procedure is carried out by means of a suitable modification of the SEM-Gibbs algorithm. The resulting methodology flexibly introduces different, and possibly user-defined, notions of cluster and, by partitioning matrices into homogeneous blocks, provides parsimonious representations of high-dimensional and complex structured time-dependent data. Lastly, the explicit modelling of time evolutions allows for meaningful interpretations of the clusters.

E1761: **Model-based clustering of directed weighted networks**
Presenter: **Volodymyr Melnykov**, The University of Alabama, United States
Co-authors: Shuchismita Sarkar, Yana Melnykov

An approach relying on the notion of mixture models is proposed for modeling and clustering directed weighted networks. The proposed methodology can be used in many settings including modeling multilayer networks. Computational issues associated with the developed procedure are addressed by the use of a MCMC procedure. An extension to modeling dynamic networks by means of mixtures of tensor normal distributions is discussed. The utility of the methodology is illustrated on synthetic data as well as real-life data containing trade operations among European Union members.

EO304 Room Virtual R21 RECENT DEVELOPMENTS FOR ELECTRONIC HEALTH DATA Chair: Mireille Schnitzer

E0675: **Use of data-adaptive analytics for high-dimensional proxy confounder adjustment in electronic healthcare databases**
Presenter: **Richard Wyss**, Brigham and Womens Hospital and Harvard Medical School, United States

Routinely-collected healthcare databases generated from insurance claims and electronic health records have tremendous potential to provide information on the real-world effectiveness and safety of medical products. However, unmeasured confounding stemming from non-randomized treatments and poorly measured comorbidities remains the greatest obstacle to utilizing these data sources for real-world evidence generation. To reduce unmeasured confounding, data-driven algorithms can be used to leverage the large volume of information in healthcare databases to identify proxy variables for confounders that are either unknown to the investigator or not directly measured in these data sources (proxy confounder adjustment). Evidence has shown that data-driven algorithms for proxy confounder adjustment can supplement investigator-specified variables to improve confounding control compared to adjustment based on investigator-specified variables alone. Consequently, there has been a recent focus on the development of data-driven methods for high-dimensional proxy confounder adjustment in healthcare database studies. We will discuss recent advancements in data-driven methods for high-dimensional proxy confounder adjustment. We will discuss challenges in assessing the validity of alternative analytic choices to tailor analyses to the given study to improve validity and robustness when estimating treatment effects in healthcare databases.

E1009: **Semi-supervised learning with electronic health records**
Presenter: **Jessica Gronsbell**, University of Toronto, United States

The adoption of electronic health records (EHRs) has generated massive amounts of routinely collected medical data with the potential to improve our understanding of healthcare delivery and disease processes. However, the analysis of EHR data remains both practically and methodologically challenging as it is recorded as a byproduct of billing and clinical care, and not for research purposes. We will discuss methods that bridge classical statistical theory and modern machine learning tools in an effort to efficiently and reliably extract insight from EHR data. We will focus primarily on (i) the challenges in obtaining annotated outcome data, such as the presence of a disease or clinical condition, from patient records and (ii) how to reduce the annotation burden by leveraging unlabeled data in model estimation and evaluation.

E0956: **Combining latent class growth models and marginal structural models to estimate the effect of treatment trajectories**
Presenter: **Denis Talbot**, Laval University, Canada
Co-authors: Awa Diop, Caroline Sirois, Jason Guertin, Bernard Candas

Latent class growth models (LCGMs) are becoming increasingly popular to summarize a time-varying treatment in a few trajectory groups. Standard approaches like a confounder-adjusted regression model or an inverse probability of trajectory groups weighted regression can yield biased estimators of the effect of the trajectory groups because some variables can have a double role of confounders and mediators. We propose to combine LCGMs with marginal structural models (MSMs) to adequately control for time-dependent confounders. The parameter of interest is defined as the projection of the true MSM onto a working model characterized by the LCGM trajectories. Inverse probability of treatment weighting, g-computation and targeted maximum likelihood estimators are proposed. These estimators are evaluated and compared with standard approaches using simulation studies. We also discuss an extension of the LCGM-MSM approach where the trajectory groups can be time-varying using a history-restricted MSM framework. The motivation was a real-world reevaluation of the efficacy and safety of statins for the primary prevention of cardiovascular disease among older adults using population-wide administrative health databases. The application of the LCGM-MSM approach in these data will be illustrated.

E1155: **Estimation of average treatment effects with binary outcomes subject to both missingness and misclassification**
Presenter: **Grace Yi**, University of Western Ontario, Canada

Causal inference has been widely conducted in various fields and many methods have been proposed for different settings. However, for noisy data with both mismeasurements and missing observations, those methods often break down. We will discuss a problem concerning estimation of the average treatment effects (ATE) when binary outcomes are subject to both missingness and misclassification. The asymptotic biases caused by ignoring missingness and/or misclassification will be examined. Methods of simultaneously correcting for missingness and misclassification effects will be discussed. Simulation studies are conducted to assess the performance of the proposed methods. An application to smoking cessation data is reported to illustrate the use of the proposed methods.

EO350 Room Virtual R22 COUNTING PROCESSES Chair: Paula Bouzas

E0291: **Clustering analysis of wildfires: A classification according to land cover types in the Valencian Community**
Presenter: **Laura Serra**, University of Girona (UdG), Spain

Typically forest fires are spread reflecting a particular pattern, which can be justified by the features of the specific place of their ignition, either from the elements of the ground that are set on fire or from their geographical location (near roads or urban centres). Within this context, the current study explains how the forest fires that occurred in the Valencian Community during the last years (from 2016 to 2020) are grouped spatially. The same work also attempts to characterise each group (cluster) in terms of location and land cover. We have performed a brief exploratory analysis of environmental covariates related to forest fires. Finite Gaussian mixture models are used to carry out the analysis. More concretely, the R package Mclust, which offers many functions to understand clustering is used. The main outcomes of this research are to offer a set of tools with which to obtain better knowledge about the type of forest fires in each studied spatial zone. Some interesting cluster patterns in certain geographical locations like river basins have also been reported. Besides, they allow measures to be applied to decrease the impacts of forest fires and help the extinction helped by the features of all forest fires grouped using spatial and land cover dimensions.

E0452: **Markov modulated Poisson processes to analyse rainfall time series**
Presenter: **Nadarajah Ramesh**, University of Greenwich, United Kingdom

Researchers in modelling rainfall time series use several stochastic point process models. Chief among them are cluster-based point process models constructed from either Bartlett-Lewis or Neyman-Scott processes. We describe a class of point process models based on a Markov modulated Poisson process that are useful in analysing rainfall time series. In particular, we discuss recent results from an exponentially decaying rainfall pulse model developed from this class of stochastic point processes. The proposed model is utilised to model hourly and sub-hourly rainfall data. The results of our analyses suggest that the proposed class of models provides a useful addition to the existing array of stochastic models for analysing fine-scale rainfall data.

E1153: **Nonparametric bias-correction and test for mark-point dependence with replicated marked point processes**
Presenter: **Ganggang Xu**, University of Miami, United States
Co-authors: Yehua Li, Yongtao Guan, Emma Jingfei Zhang

Mark-point dependence plays a critical role in research problems fitting into the general framework of marked point processes. We focus on nonparametrically adjusting for mark-point dependence when estimating the mean and covariance functions of the mark process given independent replicates of the marked point process. We assume that the mark process is a Gaussian process and the point process is a log-Gaussian Cox process, where the mark-point dependence is generated through the dependence between two latent Gaussian processes. Under this framework, naive local linear estimators ignoring the mark-point dependence can be severely biased but the biases can be corrected using a local linear estimator of the cross-covariance function. Uniform convergence rates of the bias-corrected estimators are established under mild conditions. Furthermore, we propose a formal testing procedure for mark-point dependence. The proposed test statistic, though based on nonparametric estimators, converges to an asymptotic normal distribution in a surprising parametric root-n convergence rate. The effectiveness of the proposed methods is demonstrated using extensive simulations and applications to two real data examples.

E0771: **Microstructure of foreign exchange market: Cox vs. Hawkes process**
Presenter: **Nuria Ruiz-Fuentes**, University of Jaen, Spain
Co-authors: Francisco Luguera, Paula Bouzas

The microstructure of the foreign exchange market is obviously an important field within the financial sphere; additionally, from the statistical point of view, it provides many examples of high-intensity counting processes. One such example is the number of changes of the bid and ask prices on the EUR/USD currency pair when observed in milliseconds. The Hawkes process is widely used in the literature to model this data. The intensity was calculated considering different regularization and optimization methods deriving a good model; nevertheless, the model does not provide important outcomes, such as predictions within a future interval of time. Therefore, the Cox process with intensity process modelled by FDA was considered as a worthwhile alternative to explore. In this case, having chosen the interval of time to model, less data was required to apply the estimation method, and the intensity process was stochastically estimated. Finally, the Cox process produced low error predictions and used a short processing time.

EO818 Room Virtual R23 EVALUATION OF MULTIPLE BIOMARKERS AND RELATED ROC CHARACTERISTICS Chair: Andriy Bandos

E0883: **The length of the ROC curve and the two cutoff Youden through a biomarker discovery framework**
Presenter: **Leonidas Bantis**, University of Kansas Medical Center, United States
Co-authors: John Tsimikas, Gregory Chambers, Michela Capello, Samir Hanash, Ziding Feng

During biomarker discovery, high throughput technologies allow for simultaneous input of thousands of biomarkers that attempt to discriminate between healthy and diseased subjects. In such cases, proper ranking of biomarkers is highly important. Common measures, such as the area under the receiver operating characteristic (ROC) curve (AUC), as well as affordable sensitivity and specificity levels, are often taken into consideration. Strictly speaking, such measures are appropriate under a stochastic ordering assumption, which implies that higher (or lower) measurements are more indicative of the disease. Such an assumption is not always plausible and may lead to the rejection of extremely useful biomarkers at this early discovery stage. We explore the length of a smooth ROC curve as a measure for biomarker ranking, which is not subject to a single directionality. We show that the length corresponds to a divergence, is identical to the corresponding length of the optimal (likelihood ratio) ROC curve, and is an appropriate measure for ranking biomarkers. We explore the relationship between the length measure and the AUC of the optimal ROC curve. We then provide a complete framework for the evaluation of a biomarker in terms of sensitivity and specificity through a proposed ROC analogue for use in improper settings. We consider broad parametric families as well as non-parametric estimates. We apply our approaches on real data sets related to pancreatic and esophageal cancer.

E0938: **A general framework for ROC curve inference with and without covariates**
Presenter: **Maria Xose Rodriguez-Alvarez**, BCAM, Basque Center for Applied Mathematics, Spain
Co-authors: Vanda Inacio

A general framework is proposed for the estimation of the receiver operating characteristic (ROC) curve and its conditional counterparts, the covariate-specific ROC curve and the covariate-adjusted ROC curve. The proposal builds upon the expression of the abovementioned ROC curves as the (conditional) cumulative distribution function of the so-called (conditional) placement values, i.e., the standardisation of test results in the diseased population using the non-diseased population as the reference. Estimation of the different ROC curves might therefore be approached using techniques for the estimation of (conditional) cumulative distribution functions. As an alternative, and by representing the (conditional) cumulative distribution function as the (conditional) expectation of a binary random variable, the estimation can also be carried out using a data augmentation strategy jointly with parametric or semiparametric regression methods. These different alternatives will be briefly presented and discussed, and an illustration using real data will be provided.

E0711: **A latent functional approach for modeling multi-dimensional biomarker exposures on disease risk prediction**
Presenter: **Paul Albert**, National Cancer Institute, United States
Co-authors: Sung Duk Kim

Understanding the relationships between biomarkers of exposure and disease incidence is an important problem in environmental epidemiology. Typically, a large number of these exposures are measured, and it is found either that a few exposures transmit risk or that each exposure transmits a small amount of risk, but, taken together, these may pose a substantial disease risk. Importantly, these effects can be highly non-linear and can be in different directions. We develop a latent functional approach, which assumes that the individual joint effects of each biomarker exposure can be characterized as one of a series of unobserved functions, where the number of latent functions is less than or equal to the number of exposures. We propose Bayesian methodology to fit models with a large number of exposures. An efficient Markov chain Monte Carlo sampling algorithm is developed for carrying out Bayesian inference. The deviance information criterion is used to choose an appropriate number of nonlinear latent functions. We demonstrate the good properties of the approach using simulation studies. Further, we show that complex exposure relationships can be represented with only a few latent functional curves. The proposed methodology is illustrated with an analysis of the effect of cumulative pesticide exposure on cancer risk in a large cohort of farmers.

E0200: **Doubly robust evaluation of receiver operating characteristic under covariate shift with high dimensional features**
Presenter: **Molei Liu**, Harvard T.H. Chan School of Public Health, United States

Transfer learning plays an important role in the presence of covariate shift. Most works in this regime focus on model estimation, while robust and efficient model accuracy evaluation on the target population lacks enough attention despite its importance. We tackle this problem through a novel augmented estimation approach for the receiver operating characteristic parameters. The proposed estimators are doubly robust in the sense that it is root-n consistent when one correctly specifies at least one of the two nuisance models: a density ratio model characterizing the covariate shift and an imputation model for the response Y. Our method targets a low dimensional outcome model. It accommodates high dimensional shifted features by calibrating the estimating equations for the nuisance models to correct for their estimation (regularization) bias under high dimensionality.

EO450 Room Virtual R24 REDUCTION METHODS FOR LARGE AND HIGH-DIMENSIONAL REGRESSION Chair: Katja Ickstadt

E0749: **Sparse sketches with small inversion bias**
Presenter: **Edgar Dobriban**, University of Pennsylvania, United States

For a tall $n \times d$ matrix A and a random $m \times n$ sketching matrix S, the sketched estimate of the inverse covariance matrix $(A^TA)^{-1}$ is typically biased:$E[(A_s^TA_s)^{-1}] \neq (A^TA)^{-1}$, where$A_s = SA$. This phenomenon, which we call inversion bias, arises, e.g., in statistics and distributed optimization, when averaging multiple independently constructed estimates of quantities that depend on the inverse covariance matrix. We develop a framework for analyzing inversion bias, based on our proposed concept of an (ε,δ)-unbiased estimator for random matrices. We show that when the sketching matrix S is dense and has i.i.d. sub-gaussian entries, then after simple rescaling, the estimator $(\frac{m}{m-d}A_s^TA_s)^{-1}$ is (ε,δ)-unbiased for $(A^TA)^{-1}$ with a sketch of size $m = O(d+\sqrt{d}/\varepsilon)$. In particular, this implies that for $m = O(d)$, the inversion bias of this estimator is $O(1/\sqrt{d})$, which is much smaller than the $\Theta(1)$ approximation error obtained as a consequence of the subspace embedding guarantee for sub-gaussian sketches. We then propose a new sketching technique, called LEverage Score Sparsified less-embeddings, which uses ideas from both data-oblivious sparse embeddings as well as data-aware leverage-based row sampling methods, to get ε inversion bias for sketch size $m = O(d\log d + \sqrt{d}/\varepsilon)$ in time$O(nnz(A)\log n + md^2)$, where nnz is the number of non-zeros.

E1244: **Nonuniform negative sampling and log odds correction with rare events data**
Presenter: **HaiYing Wang**, University of Connecticut, United States

The issue of parameter estimation with nonuniform negative sampling for imbalanced data is investigated. We first prove that, with imbalanced data, the available information about unknown parameters is only tied to the relatively small number of positive instances, which justifies the usage of negative sampling. However, if the negative instances are subsampled to the same level as the positive cases, there is information loss. To maintain more information, we derive the asymptotic distribution of a general inverse probability weighted (IPW) estimator and obtain the optimal sampling probability that minimizes its variance. To further improve the estimation efficiency over the IPW method, we propose a likelihood-based estimator by correcting log-odds for the sampled data and proving that the improved estimator has the smallest asymptotic variance among a large class of estimators. It is also more robust to pilot misspecification. We validate our approach on simulated data as well as a real click-through rate dataset with more than 0.3 trillion instances, collected over a period of a month. Both theoretical and empirical results demonstrate the effectiveness of our method.

E1351: **Probit regression for large data sets via coresets**
Presenter: **Christian Peters**, TU Dortmund, Germany

The purpose is to show how probit regression, an important generalized linear model for binary responses, can be solved efficiently, even when the data sets are large. To this end, we develop an algorithm that reduces a d-dimensional data set from n points to a coreset of $\text{poly}(\mu d\log n)$ weighted points, where μ is a usually small complexity parameter for compressing the data. The coreset provably allows approximating the probit loss function on the original data set up to a factor of $(1\pm\varepsilon)$ for all data sets with bounded mu-complexity, in which case the data is inseparable and the maximum likelihood estimator exists. We show how the coreset can be computed by an online algorithm that requires only one pass over the data set and $O(d^2)$ update time. The experiments on real-world data sets demonstrate that the algorithm outperforms both uniform sampling and stochastic gradient descent and that it can be applied successfully in the context of Bayesian probit regression. We also briefly discuss extensions of the standard probit model that also admit efficient sublinear approximations in the coreset framework.

E1348: **Oblivious sketching for logistic regression**
Presenter: **Simon Omlor**, TU Dortmund, Germany

What guarantees are possible for solving logistic regression in one pass over a data stream? To answer this question, we present the first data oblivious sketch for logistic regression, which is an important generalized linear model for the classification and estimation of Bernoulli probabilities. The sketching matrix can be drawn from an oblivious, i.e., data-independent distribution over sparse random matrices which is simple to implement and can be applied to a data matrix $A \in \mathbb{R}^{n\times d}$ over a turnstile data stream in input-sparsity time $O(nnz(A))$, where nnz denotes the number of non-zeros. This is important and has advantages over existing coreset constructions when it comes to high-velocity streaming applications and when data is not presented in row-order but in an arbitrary unstructured way. The resulting sketch consists of only $\text{poly}(\mu d\log n)$ weighted points, where μ is a useful parameter that captures the complexity of compressing the data. Solving (weighted) logistic regression on the sketch gives an $O(\log n)$-approximation to the original problem on the full data set. We also show how the same sketch can be slightly adapted to give an $O(1)$-approximation. Our sketches are fast, simple, easy to implement, and our experiments demonstrate that those sketching techniques are useful, practical, and competitive to uniform sampling, SGD, and to state-of-the-art coresets.

EO491 Room Virtual R25 STATISTICAL MODELING, LEARNING, AND INFERENCE Chair: Subir Ghosh

E0347: **Power analysis for knockoff-calibrated high dimensional logistic regression**
Presenter: **Jing Zhou**, KU Leuven, Belgium
Co-authors: Gerda Claeskens

Logistic regression, as a commonly used binary classification method, has been well studied in the classical setting with a fixed number of parameters p and the sample size $n \to \infty$. However, modern data structures are more versatile, allowing both $p,n \to \infty$ according to a relative growth rate. We focus on a high dimensional setting with a linear rate $p/n \to \delta \in (0,\infty)$ and a sparse coefficient vector β of which the components have a probability s to be nonzero. To estimate β, we consider the l_1-regularized logistic regression estimator $\widehat{\beta}$, of which the limiting representation is characterized by a system of equations which can be used to obtain the exact expressions of performance measures of $\widehat{\beta}$ such as the mean squared error, probability of true and false discoveries. We show that the performance measures can be used to theoretically analyze the power of the knockoff-calibrated estimators, which allows controlling the false discovery rate (FDR). Further, analytical expressions of an estimator of the FDR are derived for practical use without requiring any information of β. We evaluate the performance of the knockoff-calibrated estimators by an extensive simulation study.

E0365: **Robust fitting for generalized additive models for location, scale and shape**
Presenter: **Eva Cantoni**, University of Geneva, Switzerland
Co-authors: William Aeberhard, Giampiero Marra, Rosalba Radice

The validity of estimation and smoothing parameter selection for the wide class of generalized additive models for location, scale and shape

(GAMLSS) relies on the correct specification of a likelihood function. Deviations from such an assumption are known to mislead any likelihood-based inference and can hinder penalization schemes meant to ensure some degree of smoothness for nonlinear effects. We propose a general approach to achieve robustness in fitting GAMLSSs by limiting the contribution of observations with low log-likelihood values. Robust selection of the smoothing parameters can be carried out either by minimizing information criteria that naturally arise from the robustified likelihood or via an extended Fellner-Schall method. The latter allows for automatic smoothing parameter selection and is particularly advantageous in applications with multiple smoothing parameters. We also address the challenge of tuning robust estimators for models with nonlinear effects by proposing a novel median downweighting proportion criterion. This enables a fair comparison with existing robust estimators for the special case of generalized additive models, where our estimator competes favorably. The overall good performance of our proposal is illustrated by further simulations in the GAMLSS setting and by an application to functional magnetic resonance brain imaging using bivariate smoothing splines.

E1060: **Perturbations and causality in Gaussian latent variable models**
Presenter: **Armeen Taeb**, ETH Zurich, Switzerland

With observational data alone, causal inference is a challenging problem. The task becomes easier when having access to data from perturbing the underlying system, even when the perturbations are happening in an unspecific and non-randomized way. We provide results that enable causal discovery in this setting, and also allow for the presence of latent variables. In particular, we examine a perturbation model for interventional data over a collection of Gaussian variables. Given access to data arising from perturbations, we will introduce a regularized maximum-likelihood framework that determines the class of equally representative DAGs, and uniquely identifies the underlying causal structure under sufficiently heterogeneous data. We illustrate the effectiveness of our framework on synthetic data as well as real data involving California reservoirs.

E1185: **A simple measure conditional dependence and its application in causal inference**
Presenter: **Mona Azadkia**, ETH Zurich, Switzerland
Co-authors: Peter Buehlmann, Armeen Taeb, Sourav chatterjee

A coefficient of conditional dependence between two random variables Y and Z given a set of other variables X_1, ..., X_p, based on an i.i.d. sample is proposed. The coefficient has a long list of desirable properties, the most important of which is that under absolutely no distributional assumptions, it converges to a limit in $[0,1]$, where the limit is 0 if and only if Y and Z are conditionally independent given X_1, ..., X_p, and is 1 if and only if Y is equal to a measurable function of Z given X_1, ..., X_p. Using this statistic, we devise a new variable selection algorithm, called feature ordering by conditional independence (FOCI), which is model-free, has no tuning parameters and is provably consistent under sparsity assumptions. We focus on an application of this method in casual structure discovery.

EO481 Room Virtual R26 TOPICS IN HIGH-DIMENSIONAL STATISTICS Chair: Andreas Artemiou

E0320: **Testing for subsphericity when n and p are of different asymptotic order**
Presenter: **Joni Virta**, University of Turku, Finland

A classical test of subsphericity is extended, based on the first two moments of the eigenvalues of the sample covariance matrix, to the high-dimensional regime where the signal eigenvalues of the covariance matrix diverge to infinity and either $p/n \to 0$ or $p/n \to \infty$. In the latter case, we further require that the divergence of the eigenvalues is suitably fast in a specific sense. The developments complement earlier results in the literature that established equivalent results in the case $p/n \to \gamma \in (0,\infty)$. As a second main contribution, we use the test to derive a consistent estimator for the latent dimension of the model. Simulations and a real data example are used to demonstrate the results, also providing evidence that the test might be further extendable to a wider asymptotic regime.

E0407: **Central quantil subspace**
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 about linear and nonlinear QR models. Specifically, nonparametric estimation of the 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, we turn our attention to dimension reduction techniques for conditional quantiles and introduce a new method for reducing the dimension of the predictor X. The novelty is threefold. We start by considering a single index quantile regression model, which assumes that the conditional quantile depends on X through a single linear combination of the predictors, then extend to a multi-index quantile regression model, and finally, generalize the proposed methodology to any statistical functional of the conditional distribution. The performance of the methodology is demonstrated through simulation examples and real data applications. Our results suggest that this method has a good finite sample performance and often outperforms the existing methods.

E0454: **Unbalanced distributed estimation and inference for covariate-adjusted Gaussian graphical models**
Presenter: **Eugen Pircalabelu**, Universita catholique de Louvain, Belgium
Co-authors: Ensiyeh Nezakati Rezazadeh

A distributed estimation and statistical inference framework are introduced for the sparse precision matrix in the covariate-adjusted Gaussian graphical models under the unbalanced splitting setting. This type of splitting arises when the datasets from different sources cannot be aggregated on one single machine or when the available machines are of different powers. A de-biased estimator of the precision matrix on every single machine is proposed, and theoretical guarantees are provided. Moreover, a new de-biased estimator that is pooled across the machines using the confidence distribution is proposed. It is shown to enjoy consistency and asymptotic normality, and we provide statistical inference strategies based on it. The performance of this estimator is investigated via a simulation study and a real data example. It is shown that the performance of this estimator is close to the non-distributed estimator, which uses the entire dataset.

E0489: **Nonparametric and high-dimensional functional graphical models**
Presenter: **Eftychia Solea**, CREST and ENSAI, France
Co-authors: Holger Dette

The problem of constructing nonparametric undirected graphical models for high-dimensional functional data is considered. Most existing statistical methods in this context assume either a Gaussian distribution on the vertices or linear conditional means. We provide a more flexible model which relaxes the linearity assumption by replacing it with an arbitrary additive form. The use of functional principal components offers an estimation strategy that uses a group lasso penalty to estimate the relevant edges of the graph. We establish statistical guarantees for the resulting estimators, which can be used to prove consistency if the dimension and the number of functional principal components diverge to infinity with the sample size. We also investigate the empirical performance of our method through simulation studies and a real data application.

EO565 Room Virtual R27 SPATIAL STATISTICAL METHODS FOR MODELING EPIDEMIOLOGICAL DATA Chair: Veronica Berrocal

E0747: **Geographical evolution of lung cancer mortality by birth cohorts from 1920-29 to 1960-69 in the Italian provinces**
Presenter: **Annibale Biggeri**, University of Florence, Italy

Space-time analysis of mortality risk is useful to evaluate the epidemiological transition at the subnational level. We previously analyzed space-time variation at a small geographical scale considering as relevant time axis period or birth cohorts, and, bivariate gender-specific disease mapping.

We use almost 20 years of mortality data (1995-2016) for males and females in Italy, and we study the spatio-temporal evolution of lung cancer mortality by province and gender. We define the birth cohort as a relevant time axis. The analysis is performed using a space-time Bayesian model with space-time interaction. We found that the geographical pattern of lung cancer mortality changed during time showing that the well-known Italian north-south gradient has been replaced by a new east-west gradient. We confirmed previous results on men - where lung cancer mortality reached the peak around the birth cohort 1920-29 followed by a strong decline. We documented a different time evolution in women - where a previously undetected decline started in many Italian provinces. Different specifications are possible for the space-time interaction term: in our model interaction terms are structured both in space and time.

E0752: **Spatial co-occurrence of malignant mesothelioma and ovarian cancer in Lombardy (Italy)**
Presenter: **Dolores Catelan**, University of Padua, Italy

Multivariate disease mapping is used heuristically to summarize evidence of shared risk factors. The role of environmental asbestos in the etiology of Ovarian Cancer is widely debated. The aim is to study the spatial co-occurrence of mortality risk for Malignant Mesothelioma and Ovarian Cancer in the Lombardy region (IT) using shared Bayesian models. Malignant Mesothelioma (MM) is considered a marker of environmental asbestos exposure. Mortality data for Ovarian Cancer and MM in the Lombardy region were provided by the Italian National Health Institute for the period 2000-2018 at the municipality level. Bayesian spatial shared models were specified. Model comparison was based on Kullback-Leibler discrepancy measures and Watanabe Information criteria (WAIC). We found evidence of shared risk factors in some areas of the Lombardy region, while in others the results are dependent on the prior model specifications.

E1013: **How close and how much: Linking health outcomes to spatial distributions of built environment features**
Presenter: **Veronica Berrocal**, University of California, Irvine, United States

Built environment features (BEFs) refer to aspects of the human-constructed environment, which may in turn support or restrict health-related behaviors and thus impact health. We are interested in understanding whether the spatial distribution and quantity of fast-food restaurants (FFRs) influence the risk of obesity in schoolchildren. To achieve this goal, we propose a two-stage Bayesian hierarchical modeling framework. In the first stage, examining the position of FFRs relative to that of some reference locations - in our case, schools - we model the distances of FFRs from these reference locations as realizations of Inhomogenous Poisson processes (IPP). With the goal of identifying representative spatial patterns of exposure to FFRs, we model the intensity functions of the IPPs using a Bayesian non-parametric view and specifying a Nested Dirichlet Process prior. The second stage model relates exposure patterns to obesity, offering two different approaches to accommodate uncertainty in the exposure patterns estimated in the first stage. Our analysis on the influence of patterns of FFR occurrence on obesity among Californian schoolchildren has indicated that, in 2010, among schools that are consistently assigned to a cluster, there is a lower odds of obesity amongst 9th graders who attend schools with most distant FFR occurrences in a 1-mile radius as compared to others.

E1019: **Heterogeneous effects of the built environment**
Presenter: **Brisa Sanchez**, Drexel University, United States

An approach is presented to estimate distance-dependent heterogeneous associations between point-referenced exposures to built environment characteristics and health outcomes. By estimating associations that depend non-linearly on the distance between subjects and point-referenced exposures, this method addresses the modifiable area-unit problem that is pervasive in the built environment literature. Additionally, by estimating heterogeneous effects, the method also addresses the uncertain geographic context problem. The key innovation of our method is to combine ideas from the non-parametric function estimation literature and the Bayesian Dirichlet process literature. The former is used to estimate nonlinear associations between the subject's outcomes and proximate built environment features, and the latter identifies clusters within the population that have different effects. We study this method in simulations and apply our model to study heterogeneity in the association between fast-food restaurant availability and weight status of children attending schools in Los Angeles, California.

EO671 Room Virtual R28 STATISTICAL METHODS FOR DATA INTEGRATION IN BIOMEDICAL RESEARCH Chair: Rui Duan

E0398: **Survival analysis in multi-site studies using summary-level risk set tables**
Presenter: **Di Shu**, University of Pennsylvania and Children's Hospital of Philadelphia, United States

Medical research often analyzes data from multiple sources to increase statistical power and generalizability. A growing number of studies are now conducted within multi-site distributed data networks. For example, the Sentinel System, funded by the U.S. Food and Drug Administration, monitors the safety of approved medical products using data from multiple data partners. Within these networks like the Sentinel System, each data partner maintains physical control of their data and may not always be able or willing to share individual-level data for analysis. We will introduce a one-step method that allows data partners to share only summary-level risk set tables to estimate overall and site-specific hazard ratios. We will also discuss how to apply risk set tables to other important measures such as Kaplan-Meier curves, as well as some future topics. We will justify the method theoretically, illustrate its use, and demonstrate its statistical performance using both real-world and simulated data.

E0304: **A framework for data integration with dependence and heterogeneity**
Presenter: **Emily Hector**, North Carolina State University, United States
Co-authors: Peter Song

A framework is proposed to jointly estimate regression parameters from multiple, potentially heterogeneous data sources with correlated vector outcomes. The primary goal of this joint integrative analysis is to estimate covariate effects on all outcomes through a marginal regression model in a statistically and computationally efficient way. We develop a data integration procedure for statistical estimation and inference of regression parameters that is implemented in a fully distributed and parallelized computational scheme. To overcome computational and modeling challenges arising from the high-dimensional likelihood of the correlated vector outcomes, we propose to analyze each data source using quadratic inference functions, and then to jointly reestimate parameters from each data source by accounting for correlation between data sources using a combined meta-estimator in a similar spirit to the generalized method of moments. We show both theoretically and numerically that the proposed method yields efficiency improvements and is computationally fast. We illustrate the proposed methodology with the joint integrative analysis of the association between smoking and metabolites in a large multi-cohort study.

E0722: **Communicating likelihood profiles of single parameters of interest in complex models in federated research networks**
Presenter: **Martijn Schuemie**, Janssen Research and Development, United States

Effects of medical intervention are increasingly studied in distributed research settings, using multiple clinical data sources such as electronic health records and administrative claims. Sharing individual patient data is seldom allowed, and instead, only summary statistics can be used to combine evidence across the network. Although the models in these studies can be complex, for example, Cox models conditioned on propensity score strata, or multi-variable Poisson models conditioned on individual patients, there is typically only a single parameter of interest, representing the effect of the exposure on the outcome. We present likelihood profiles as a generic approach to communicating information on this parameter between sites that is both privacy-preserving and efficient, requiring only a single round of communication. A likelihood profile is a simple representation of the likelihood of values of the parameter over a wide range, for example by sampling the likelihood on a grid and using linear interpolation. We demonstrate how this approach can be used in both fixed and (Bayesian) random-effects models, using both real and simulated data. Results show performance comparable to pooling data.

E0332: **Targeting underrepresented populations in precision medicine: A federated transfer learning approach**
Presenter: **Rui Duan**, Harvard University, United States

One serious challenge in precision medicine research is the limited representation of minorities and disadvantaged populations, such as populations with low socioeconomic status and racial and ethnic minority groups. To advance prediction medicine, it is crucial to improve the performance of statistical and machine learning models in underrepresented populations so as not to exacerbate health disparities. We address the lack of representation and disparities in model performance through two strategies: (1) leverage the shared knowledge from diverse populations, and (2) integrate larger bodies of data from multiple healthcare organizations. More specifically, we develop transfer learning strategies to transfer the shared knowledge learned from diverse populations to an underrepresented population, so that comparable model performance can be reached with much fewer data. On the other hand, we propose federated learning methods to increase the sample sizes of underrepresented populations and the diversity of the data through multi-center collaborative research via a safe and efficient way. Our methods have solid theoretical foundations. We demonstrate the feasibility and validity of our methods through numerical experiments and a real application to a multi-center study for constructing polygenic risk prediction models for Type II Diabetes.

EO790 Room Virtual R29 NEW ADVANCEMENTS IN SEMIPARAMETRIC AND NONPARAMETRIC METHODS Chair: Pengfei Li

E0298: **Bayesian jackknife empirical likelihood**
Presenter: **Yichuan Zhao**, Georgia State University, United States
Co-authors: Yichen Cheng

The empirical likelihood is a very powerful nonparametric tool that does not require any distributional assumptions. It has been shown that if you replace the usual likelihood component in the Bayesian posterior likelihood with the empirical likelihood, then the posterior inference is still valid when the functional of interest is a smooth function of the posterior mean. However, it is not clear whether similar conclusions can be obtained for parameters defined in terms of U-statistics. We propose the so-called Bayesian jackknife empirical likelihood, which replaces the likelihood component with the jackknife empirical likelihood. We show, both theoretically and empirically, the validity of the proposed method as a general tool for Bayesian inference. Empirical analysis shows the small sample performance of the proposed method is better than its frequentist counterpart. Analysis of a case-control study for pancreatic cancer is used to illustrate the new approach.

E0306: **PAVA-assisted learning with application on estimating optimal individualized treatment regimes**
Presenter: **Baojiang Chen**, University of Texas Health Science Center at Houston – Austin Regional Campus, United States
Co-authors: Ao Yuan, Jing Qin

Personalized medicine allows individuals to choose the best fit of their treatments based on their characteristics through an individualized treatment regime. We develop a pool adjacent violators algorithm-assisted learning method to find the optimal individualized treatment regime under the monotone single index outcome gain model. The proposed estimator is more efficient than peers, and it is robust to the misspecification of the propensity score model or the baseline regression model. The optimal treatment regime is also robust to the misspecification of the functional form of the expected outcome gain model. Simulation studies verified our theoretical results. We also provide an estimate of the expected outcome gain model. Plotting the expected outcome gain versus an individual's characteristics index can visualize how significant the treatment effect is over the control. We apply the proposed method to an AIDS study.

E1118: **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 distribution of the random error is completely unknown; existing methods either need strong assumptions, or are less effective when the distribution of the random error significantly deviates from the normal distribution. We adopt the semiparametric assumption and propose a maximum profile binomial likelihood method. We theoretically establish the joint distribution of the estimators of the model parameters. Through extensive numerical studies, we demonstrate that our method has an advantage over existing methods, especially when the distribution of the random error deviates from the normal distribution. Furthermore, we compare the performance of our method and existing methods on an HIV data set.

E1120: **Semiparametric inference on Gini indices of two semicontinuous populations under density ratio models**
Presenter: **Meng Yuan**, University of Waterloo, Canada
Co-authors: Pengfei Li, Changbao Wu

The Gini index is a popular inequality measure with many applications in social and economic studies. The focus is on semiparametric inference on the Gini indices of two semicontinuous populations. We characterize the distribution of each semicontinuous population by a mixture of a discrete point mass at zero and a continuous skewed positive component. A semiparametric density ratio model is then employed to link the positive components of the two distributions. We propose the maximum empirical likelihood estimators of the two Gini indices and their difference, and further investigate the asymptotic properties of the proposed estimators. The asymptotic results enable us to construct confidence intervals and perform hypothesis tests for the two Gini indices and their difference. We show that the proposed estimators are more efficient than the existing fully nonparametric estimators. The proposed estimators and the asymptotic results are also applicable to cases without excessive zero values. Simulation studies show the superiority of our proposed method over existing methods.

EO774 Room Virtual R30 STATISTICS FOR SPDES Chair: Mathias Trabs

E0431: **Nonparametric calibration for stochastic reaction-diffusion equations based on discrete observations**
Presenter: **Florian Hildebrandt**, University of Hamburg, Germany
Co-authors: Mathias Trabs

In view of a growing number of stochastic partial differential equation (SPDE) models used in the natural sciences and mathematical finance, their data-based calibration has become an increasingly active field of research during the last few years. Nonparametric estimation for semilinear SPDEs, namely stochastic reaction-diffusion equations in one space dimension, is discussed. We consider observations of the solution field on a discrete grid in time and space with infill asymptotics in both coordinates. Firstly, based on a precise analysis of the Hoelder regularity of the solution process and its nonlinear component, we deduce that the asymptotic properties of diffusivity and volatility estimators derived from realized quadratic space-time variations in the linear setup generalize to the semilinear SPDE. Doing so, we obtain a rate-optimal joint estimator of the two parameters. Secondly, we present a nonparametric estimator for the reaction function specifying the underlying equation. The estimate is chosen from a finite-dimensional function space based on a least-squares criterion. An oracle inequality with respect to the L^2-risk provides conditions for the estimator to achieve the usual nonparametric convergence rate. Adaptivity is provided via model selection.

E0592: **A power variation approach to statistical analysis of discretely sampled semilinear SPDEs**
Presenter: **Igor Cialenco**, Illinois Institute of Technology, United States

Motivated by problems from statistical analysis for discretely sampled SPDEs, we derive central limit theorems for higher-order finite differences

applied to stochastic processes with arbitrary finitely regular paths. We prove a new central limit theorem for some power variations of the iterated integrals of a fractional Brownian motion (fBm) and consequently apply them to the estimation of the drift and volatility coefficients of semilinear stochastic partial differential equations driven by an additive Gaussian noise white in time and possibly colored in space. In particular, we show that approximating naively derivatives by finite differences in certain estimators may introduce a nontrivial bias that we compute explicitly.

E1333: **High-frequency analysis of parabolic stochastic PDEs**
Presenter: **Carsten Chong**, Columbia University, United States
The focus is on the stochastic heat equation driven by an additive or multiplicative Gaussian noise that is white in time and spatially homogeneous in space. Assuming that the spatial correlation function is given by a Riesz kernel of order α, we prove a central limit theorem for the power variations of the solution in the additive case. We further show that the same central limit theorem is valid with multiplicative noise if $\alpha \in (0,1)$ but fails in general if $\alpha = 1$ (and $d \geq 2$) or if the noise is a space-time white noise (and $d = 1$). We discuss our results in the context of statistical estimation for the stochastic heat equation.

E1400: **Parameter estimation for anisotropic SPDEs from multiple local measurements**
Presenter: **Randolf Altmeyer**, Cambridge University, United Kingdom
Co-authors: Martin Wahl, Anton Tiepner
The purpose is to discuss how the coefficients in a general second-order linear stochastic partial differential equation (SPDE) can be estimated. Given multiple spatially localised measurements, estimators for the diffusivity, transport and reaction coefficients are constructed. All coefficients are allowed to be anisotropic. With the spatial resolution of the measurements tending to zero and assuming a growing number of measurements, the constructed estimators achieve the optimal rate of convergence. For each coefficient, this rate depends on the order of the respective differential operator with the best rate for the diffusivity and the worst rate for the reaction terms.

EO625 Room Virtual R32 NOVEL PERSPECTIVES IN BAYESIAN STATISTICS Chair: Pier Giovanni Bissiri

E0492: **Bayesian uncertainty**
Presenter: **Stephen Walker**, University of Texas at Austin, United States
Co-authors: Chris Holmes, Edwin Fong
It is argued that quantifying Bayesian uncertainty is concerned with placing a distribution on the missing data, which, if known, the parameter of interest would be fully known. Hence, the missing data arise from the usual observations being a finite sample. Under certain assumptions on the distribution for the missing data, it is possible to recover the usual posterior distribution. With alternative assumptions, e.g. replacing exchangeability for conditionally identically distributed, we derive "posterior" distributions obtained as limits of martingale sequences. Illustrations will be presented.

E0329: **Asymptotic concentration of Gibbs posterior distributions**
Presenter: **Nicholas Syring**, Iowa State University, United States
Bayesian posterior distributions are widely used for inference, but their dependence on a statistical model creates some challenges. In particular, there may be lots of nuisance parameters that require prior distributions and posterior computations, plus a potentially serious risk of model misspecification bias. Gibbs posterior distributions, on the other hand, offer direct, principled, probabilistic inference on quantities of interest through a loss function, not a model-based likelihood. Here we provide simple sufficient conditions for establishing Gibbs posterior concentration rates when the loss function is of a sub-exponential type. We apply these general results in a range of practically relevant examples, including mean regression, quantile regression, and sparse high-dimensional classification. We also apply these techniques in an important problem in medical statistics, namely, estimation of a personalized minimum clinically important difference.

E0792: **Fast learning rate selection for Bayesian non-parametric quantile regression**
Presenter: **Matteo Fasiolo**, University of Bristol, United Kingdom
Quantile regression (QR) models are often fitted to data by minimising the so-called check or pinball loss. In a Bayesian framework, quantile regression can be based on the asymmetric Laplace (AL) distribution, because the resulting negative log-likelihood corresponds to the pinball loss. In a non-parametric spline smoothing context, it is tempting to use (a smooth generalization of) the AL distribution in conjunction with standard likelihood-based methods to fit non-parametric QR models. We will explain that this leads to poor results both in terms of smoothness of the fit and of frequentist coverage of the resulting credible intervals. We will also discuss how the issue can be alleviated via a calibration step aimed at efficiently selecting the learning rate balancing the relative weights of the loss based likelihood and of the smoothing priors.

E0653: **An objective prior from a scoring rule**
Presenter: **Cristiano Villa**, Newcastle University, United Kingdom
Co-authors: Stephen Walker
A novel objective prior distribution levering on the connections between information, divergence and scoring rules, is introduced. In particular, we do so from the starting point of convex functions representing information in density functions. This provides a natural route to proper local scoring rules using Bregman divergence. Specifically, we determine the prior which solves setting the score function to be a constant. Although in itself this provides motivation for an objective prior, the prior also minimizes a corresponding information criterion.

EO422 Room Virtual R37 STATISTICAL METHODS FOR MULTI-MODAL IMAGING DATA Chair: Kristin Linn

E0706: **Two sample testing for diffusion tensor imaging data**
Presenter: **Gina-Maria Pomann**, Duke University, United States
Co-authors: Ana-Maria Staicu, Sujit Ghosh
Motivated by a natural history imaging study, we present a non-parametric testing procedure for testing the null hypothesis that two samples of curves observed at discrete grids and with noise have the same underlying distribution. We use functional principal components-based methods to develop a test for the equality of the distributions of two samples of curves, when their eigenfunctions are the same. The approach reduces the dimensionality of the testing problem in a way that enables the application of traditional nonparametric univariate testing procedures. This results in a procedure that is not only computationally efficient but also allows for a variety of sampling designs. This methodology is applied to a diffusion tensor imaging (DTI) study, where the objective is to statistically compare white matter tract profiles between healthy individuals and multiple sclerosis patients, as assessed by conventional DTI measures.

E0257: **Graph-theoretic modeling of brain functional connectivity**
Presenter: **Ani Eloyan**, Brown University, United States
Functional connectivity (FC) has been used to study functional associations among pairs of brain regions and identify temporal correlations between neurophysiological events. FC is estimated using data collected by functional magnetic resonance imaging technology. We consider the estimation of task FC during a motor task. Data are publicly available from the Human Connectome Project. One of the approaches for estimation of FC is the implementation of graph-theoretic methods. Since FC refers to the estimation of undirected temporal associations between any two regions in the brain, often including spatially incongruous areas, a graph with vertices corresponding to brain regions of interest and edges corresponding to existing connections between regions is used as a model for FC. We will review various approaches to FC estimation using graph-theoretic

methods and propose a novel estimation procedure incorporating structural connectivity estimated by diffusion tensor imaging. We will discuss the comparisons of the proposed approach with other methods for the estimation of FC and their computational efficiency.

E0372: **Quantitative susceptibility maps in multiple sclerosis lesions**
Presenter: **Elizabeth Sweeney**, Weill Cornell, United States

Multiple sclerosis (MS) is an inflammatory disease of the central nervous system characterized by lesions in the brain and spinal cord. Magnetic resonance images (MRI) are sensitive to these lesions. A particular type of lesion, called a chronic active lesion, is characterized by a hyperintense rim of iron-enriched, activated microglia and macrophages, and has been linked to greater tissue damage. An MRI technique called quantitative susceptibility mapping (QSM) provides efficient in vivo quantification of susceptibility changes related to iron deposition and identifies these chronic active lesions, called QSM rim positive (rim+) lesions. QSM rim+ MS lesions and their longitudinal behavior have the potential to serve as a biomarker of chronic inflammation and to be utilized to monitor disease progression and evaluate disease-modifying therapies in MS. We will discuss the challenges of estimating treatment effects using the longitudinal behavior of QSM rim+ lesions. We will compare two disease-modifying treatments, Tecfidera and Copaxone, using linear mixed-effects regression models with inverse probability of censor weighting. One of the major limitations of this model is that the inflammatory stage or age of the lesion is unknown, causing misregistration of the lesion-level data. We will also introduce a methodology to estimate the age of MS lesions using both cross-sectional and longitudinal MRI information.

E0227: **Multiple sclerosis diagnostics via statistical analysis of MRI**
Presenter: **Russell Shinohara**, University of Pennsylvania, United States

Lesions in the white matter of the brain, including those that arise in multiple sclerosis, are abnormalities measurable on MRI. While much literature has focused on identifying these lesions, less work has focused on the nature of these lesions. As new imaging modalities arise that allow us to interrogate these lesions better, new statistical modeling problems that include spatial constraints and overlapping analysis domains are increasingly important. Leveraging multi-modal imaging approaches that focus on knowledge about etiology is critical for developing the next generation of robust and generalizable diagnostic imaging biomarkers.

EO742 Room Virtual R38 RECENT DEVELOPMENTS ON MEDIATION AND PATH ANALYSIS Chair: Marco Doretti

E0402: **Causal mediation analysis with double machine learning**
Presenter: **Martin Huber**, University of Fribourg, Switzerland
Co-authors: Helmut Farbmacher, Lukas Laffers, Henrika Langen, Martin Spindler

Causal mediation analysis is combined with double machine learning to control for observed confounders in a data-driven way under a selection-on-observables assumption in a high-dimensional setting. We consider the average indirect effect of a binary treatment operating through an intermediate variable (or mediator) on the causal path between the treatment and the outcome, as well as the unmediated direct effect. Estimation is based on efficient score functions, which possess a multiple robustness property w.r.t. misspecifications of the outcome, mediator, and treatment models. This property is key for selecting these models by double machine learning, which is combined with data splitting to prevent overfitting in estimating the effects of interest. We demonstrate that the direct and indirect effect estimators are asymptotically normal and root-n-consistent under specific regularity conditions and investigate the finite sample properties of the suggested methods in a simulation study when considering lasso as a machine learner. We also provide an empirical application to the U.S. National Longitudinal Survey of Youth, assessing the indirect effect of health insurance coverage on general health operating via routine checkups as a mediator, as well as the direct effect. We find a moderate short-term effect of health insurance coverage on general health, which is, however, not mediated by routine checkups.

E0854: **Path weights in concentration graph models**
Presenter: **Alberto Roverato**, University of Padova, Italy

Statistical models associated with a graph, called graphical models, are of significant interest in many modern applications and have become a popular tool for representing network structures in applied contexts such as genetic and brain network analysis. A graph is represented as a set of nodes, also called vertices, interconnected by a set of edges. A path in a graph is a finite sequence of distinct vertices with the property that each vertex in the sequence is adjacent to the vertex next to it. In graphical models, paths joining vertices of the graph play a central role because they determine the specification of the association structure of the variables highlighting the role played by intermediate variables. In models for continuous acyclic directed graphs, the well-established theory of path analysis provides a method that aims at quantifying the relative importance of causal relationships represented by directed paths. On the other hand, in undirected graphs a theory concerning the analysis of the strength of the association encoded by paths has been introduced only more recently. We consider concentration graph models and show how weights associated with undirected paths can be applied in the analysis of the graph structure and in the computation of betweenness centrality measures.

E0916: **Mediation in case control studies when both the outcome and mediator are binary**
Presenter: **Minna Genback**, Umea University, Sweden
Co-authors: Marco Doretti, Elena Stanghellini

Given a treatment X, a mediator M and an outcome Y, the aim of mediation analysis is to decompose the total (marginal) effect of X on Y into a direct one and an indirect one, i.e. mediated by M. This decomposition is usually made in a nonparametric way, though parametric contributions also exist. Parametric contributions are of particular interest when considering continuous treatments. We focus on a situation with both M and Y are binary, and we assume that they can be modelled via a series of univariate logistic regressions, possibly with covariates and the interaction term between M and X in the outcome equation. When the outcome is rare, in order to increase precision, it is customary to perform outcome dependent sampling designs, such as case-control studies. In this context, although features of the conditional distribution of Y given X and M can be identified, also nonparametrically, the conditional distribution of M given X, which is necessary to perform mediation analysis, is distorted. We here present a procedure to identify and estimate the parameters (by M estimation and maximum likelihood) of the logistic models of interest and therefore to perform mediation analysis.

E1534: **Interventional effects with latent mediators: Applications in life course epidemiology**
Presenter: **Bianca Lucia De Stavola**, University College London GOS Institute of Child Health, United Kingdom

Life-course epidemiology is important for understanding the effects of lifetime exposures on health outcomes, and for suggesting interventions to improve health, but it is complex and challenging. Conceptual models, represented by formal directed acyclic graphs, can help make sense of the many possible pathways via which exposure in early life may affect health outcomes in adulthood, while counterfactual reasoning can help specify relevant targets of estimation. Using an example drawn from life course epidemiology, alternative conceptual models will be discussed to represent the role that body size at increasing ages in childhood may play in the onset of eating disorders in adolescence. Different causal questions can be asked and expressed for example in terms of natural effects of early childhood size mediated or not mediated via later childhood sizes. These investigations will then be expanded to address questions regarding the best time for intervening to prevent eating disorders. Estimands such as the direct and indirect interventional effects will be discussed, involving a single or multiple mediators, as well as observed or latent mediators.

EO515 Room Virtual R39 ADVANCES IN LONGITUDINAL DATA MODELLING Chair: Maria Francesca Marino

E0588: **Variable selection in hidden Markov models with missing data**
Presenter: **Fulvia Pennoni**, University of Milano-Bicocca, Italy

Co-authors: Francesco Bartolucci, Silvia Pandolfi

A novel variable and model selection method is proposed to analyze multiple time-series and panel data based on a Hidden Markov (HM) model for multivariate continuous responses. We consider an approach for inference, under the missing-at-random assumption to account for missing data, focusing on the maximum likelihood estimation of the model parameters through a modified Expectation-Maximization (EM) algorithm. We develop a greedy forward-backwards algorithm based on the Bayesian Information Criterion (BIC) seen as an approximation of the Bayes factor. In this way, we achieve a dimensionality reduction of the complete set of response variables to a smaller subset and thus we select the most useful variables for clustering purposes. The BIC is also used to choose the optimal number of latent states during the steps of the greedy search algorithm. In applying the selection method, the estimation of multivariate linear regression models is required. In the presence of missing values in the set of independent variables, we adopt a sort of multiple imputations based on the posterior expected values obtained at the convergence of the EM algorithm of the estimated HM model. To illustrate the proposal we use a collection of macroeconomic indicators provided by the World Bank related to 217 countries followed over a long period of time. The chosen HM model allows us to dynamically characterize countries' transitions between hidden states representing different levels of development.

E0799: **Divisive hierarchical bayesian clustering of longitudinal data**

Presenter: **Daniel Sewell**, University of Iowa, United States

Co-authors: Elliot Burghardt, Joseph Cavanaugh

Discovering hidden subgroups of a population with differing temporal trends is critical in clinical medicine, allowing researchers to design more powerful studies, personalize treatment options, and provide more information to patients about their condition. Clustering longitudinal data to find these subgroups is challenging due to temporal dependence, differing number of observations per subject, and the need to model multivariate trajectories. We propose a model-based clustering algorithm that takes advantage of the relationships between observations across variables and across time. We use a divisive hierarchical method which provides guidance on the plausible number of clusters in a principled way while still allowing for scientific insight to guide the selection process, and yields at each level of the hierarchy a valid estimate of the partition of the data. Our methods were inspired by the need to find subgroups in patients with Parkinson's Disease based on disease progression. We applied our method to the Parkinson's Progression Markers Initiative and discovered meaningful subgroups with varying rapidity of progression which is able to be predicted using baseline data.

E0984: **Handling endogeneity in linear quantile regression models for longitudinal data**

Presenter: **Marco Alfo**, University La Sapienza, Rome, Italy

Co-authors: Francesca Martella

Individual specific effects are often included in regression models for longitudinal data. These are used to account for the effect of time-constant unobserved covariates, often referred to as unobserved heterogeneity, which may be considered as random variables (random effects). In this case, the individual-specific effects may also account for inter-individual dependence. However, the assumptions on the dependence between the random effects and the observed covariates, that is between unobserved and observed covariates, are a crucial point when integrating out the random effects to derive the model likelihood. If dependence is not properly taken into account, the resulting estimator may be inconsistent. There exist solutions in the linear regression context and extensions to the context of generalized linear models, the correlated random effect estimator. Its properties are due to the geometric properties of the (generalized) least-squares method. When we move to linear quantile regression, the same properties are not valid any longer and this estimator may not be optimal. We review methods to deal with endogeneity in the context of mixed linear quantile regression for longitudinal data and propose a general solution by exploiting a finite mixture specification.

E0255: **Latent Ornstein-Uhlenbeck models for Bayesian analysis of multivariate longitudinal categorical responses**

Presenter: **Emmanuel Lesaffre**, University of Leuven, Belgium

Co-authors: Trung Dung Tran

To explore the association of oral health with general health information obtained from a registry done on the elderly population in Belgium, we propose a Bayesian latent vector autoregressive (LVAR) model. This model handles multivariate balanced longitudinal data of binary and ordinal variables (items) as a function of a small number of continuous latent variables. We focus on the evolution of the latent variables while taking into account the correlation structure of the responses. Often local independence is assumed. Local independence implies that, given the latent variables, the responses are assumed mutually independent cross-sectionally and longitudinally. However, conditioning on the latent variables may not remove the dependence of the responses. We address local dependence by further conditioning on item-specific random effects. In a second step we extend the previous model to the unbalanced case. This model is then generalized to analyse multivariate unbalanced longitudinal data. We show that assuming real eigenvalues for the drift matrix of the OU process, as is frequently done, can lead to biased estimates and/or misleading inference when the true process is oscillating. Our proposal allows for both real and complex eigenvalues. We illustrate our model with a dataset containing patients with amyotrophic lateral sclerosis disease. We were interested in how bulbar, cervical, and lumbar functions evolve over time. Simulations showed a good behaviour.

EO621 Room Virtual R40 STATISTICAL LEARNING AND REGULARIZED REGRESSION Chair: Thomas Kneib

E0503: **A hybrid machine learning approach for the modeling and prediction of the UEFA EURO2020**

Presenter: **Andreas Groll**, Technical University Dortmund, Germany

Conventional approaches that analyze and predict the results of international matches in football are mostly based on the framework of Generalized Linear Models. The most frequently used type of regression model in the literature is the Poisson model. It has been shown that the predictive performance of such models can be improved by combining them with different regularization methods such as penalization. More recently, also methods from the machine learning field such as boosting and random forests turned out to be very powerful in the prediction of football match outcomes. We analyze both a hybrid random forest extension based on conditional inference trees and a hybrid boosting extension based on extreme gradient boosting for modeling football matches. The models are fitted to match data from previous UEFA European Championships (EUROs) and based on the corresponding estimates all match outcomes of the EURO 2020 are repeatedly simulated (100,000 times), resulting in winning probabilities for all participating national teams.

E0725: **Variable selection and allocation in joint models via gradient boosting techniques**

Presenter: **Colin Griesbach**, Georg-August-University Goettingen, Germany

Co-authors: Andreas Mayr, Elisabeth Bergherr

Modelling longitudinal data and risk for events separately, even though the underlying processes are related to each other, leads to loss of information and bias. Hence, the popularity of joint models for longitudinal and time-to-event data has grown rapidly in the last few decades. Gradient boosting is a statistical learning method that has the inherent ability to select variables and estimate them at the same time. We construct a data-driven allocation algorithm for basic joint models by applying gradient boosting. Instead of specifying beforehand which covariate has an influence on which part of the joint model, the algorithm allocates the covariates to the appropriate sub-model. A simulation study shows that this is possible when using a non-cyclic updating scheme for the boosting algorithm. In addition, recent findings of adaptive step lengths and early stopping based on probing are incorporated in order to improve allocation accuracy and reduce the computational effort.

E0890: **LIESEL: A software framework for prototyping Bayesian models and exploring estimation methods**

Presenter: **Paul Wiemann**, TU Dortmund University, Germany

Co-authors: Hannes Riebl, Thomas Kneib

LIESEL is a software framework for Python to facilitate statistical research on Bayesian models focusing on modularity, extensibility, and reliability. In the area of statistical software, LIESEL is located between specialized implementations of specific model classes and general-purpose software packages for Bayesian inference. The framework can be used for quick model prototyping or serves as a basis for the implementation of complex models or novel statistical inference algorithms. High-performance, albeit being implemented in Python, is achieved by the extensive use of the modern libraries JAX and TensorFlow Probability providing access to just-in-time compilation, automatic differentiation, and vectorization. Furthermore, LIESEL runs on high-performance computing devices like GPUs or TPUs. The framework provides an easily extensible library for MCMC estimation, including the HMC and NUTS sampling algorithms. Because of the strong modularity, LIESEL's estimation and modeling modules can be used independently of each other. LIESEL comes with tools to set up structured additive distributional regression models. Distributional regression enables researchers to explore complex relationships between explanatory and response variables beyond the mean. We present LIESEL's architecture and demonstrate its applicability in several case studies featuring distributional regression, including copula regression.

E1111: **Scalable distributional learning**

Presenter: **Nikolaus Umlauf**, University of Innsbruck, Austria

Estimating distributional regression models with very large data sets is a difficult task. In particular, the use of non-standard distributions can easily lead to memory-related but also efficiency problems, which usually results in the models not being able to be estimated at all, sometimes even on high-performance computers. We, therefore, propose a novel backfitting algorithm that is based on the ideas of stochastic gradient descent and can deal virtually with any amount of data on a conventional laptop. Moreover, the algorithm performs automatic variable and smoothing parameter selection and its performance is in most cases superior or at least equal to other implementations for distributional regression. With this new algorithm, we demonstrate the estimation of complex distribution regression models with a challenging example using a new, very large dataset on child undernutrition in low- and middle-income countries.

EO615 Room K2.40 (Hybrid 08) ADVANCES IN THE STATISTICAL ANALYSIS OF NEUROIMAGING DATA Chair: John Kornak

E0342: **Scalable Bayesian models for inference of networks and covariate effects**

Presenter: **Marina Vannucci**, Rice University, United States

New methods for the simultaneous inference of graphical models and covariates effects in the Bayesian framework will be discussed. We will consider settings where we are interested in the estimation of sparse networks among a set of primary variables, where covariates may impact the strength of edges. The proposed model utilizes spike-and-slab priors to perform edge selection, and Gaussian process priors to allow for flexibility in the covariate effects. In order to estimate these models, we rely on efficient deterministic algorithms based on variational inference. Simulation studies demonstrate how the proposed model improves on the accuracy of previous models in both network recovery and covariate selection. We apply the proposed model to fMRI data, learning both a functional network between brain regions and how the strength of network edges varies based on subject-level covariates, such as age and gender.

E0707: **Bayesian inferences on neural activity in EEG-based brain-computer interface**

Presenter: **Jian Kang**, University of Michigan, United States

A brain-computer interface (BCI) is a system that uses brain activity to control or communicate with technology. In particular, BCIs help people with disabilities use technology for communication. A common design for an electroencephalogram (EEG) BCI relies on the classification of the P300 event-related potential (ERP), which is a response elicited by the rare occurrence of target stimuli among common non-target stimuli. The existing studies have focused on constructing the ERP classifiers, but few provide insights into the underlying mechanism of the neural activity. To this end, we perform a novel Bayesian analysis of the probability distribution of the multi-channel EEG signals from real participants under the P300 ERP-BCI design. In contrast to classification, our goal is to identify relevant spatial-temporal differences of the neural activity in response to different external stimuli, which provides statistical evidence of P300 waveforms and facilitates designing user-specific profiles for efficient brain-computer communications. We also perform sensitivity and reproducibility analyses, make cross-participant comparisons, and design simulation studies to show the robustness of our analysis.

E1007: **Mass univariate modelling for binary-valued neuroimaging data**

Presenter: **Thomas Nichols**, University of Oxford, United Kingdom

Co-authors: Petya Kindalova, Ioannis Kosmidis

While the vast majority of brain image data are continuous, there is growing interest in binary-valued images describing brain lesions in Magnetic Resonance Imaging (MRI). Binary image data can identify the tissue damaged by a stroke infarct, Multiple Sclerosis lesions in white matter, or bright spots simply called white matter hyperintensities (WMH). While various sophisticated analyses can be proposed, a basic mass univariate regression is vital to map out the influence of the explanatory variables in an unbiased manner. However, the base rate of lesion incident is often very low, leading to many voxels with total- or quasi-separation. We propose a comprehensive simulation framework to evaluate methods for this type of data. We use WMH data from the UK Biobank ($N = 40,000$) to define a realistic simulation model that accounts for the spatial dependence in the lesions. Generating arbitrary-N data with covariates where ground truth is known, we compare 3 probit regression methods: Maximum Likelihood (ML), Penalised ML (PML) with Jeffreys prior (aka Firth regression), and a Bayesian Spatial Generalized Linear Mixed Model previously proposed. We find that ML often has problems with separation that PML avoids, and for the smallest sample sizes, the Bayesian method has the best MSE. We will discuss extensions for repeated measures mass univariate modelling with penalised GEE.

E1609: **Bayesian lesion estimation with a structured spike-and-slab prior**

Presenter: **Habib Ganjgahi**, Statistics Department, University of Oxford, United Kingdom

Co-authors: Anna Menacher, Chris Holmes, Thomas Nichols

Neural demyelination and damages to the human brain nervous system appear at hyperintense areas in magnetic resonance imaging (MRI) scans, known as lesions. Modelling binary images at the population level, where each voxel represents the existence of a lesion, plays an important role in understanding ageing and inflammatory diseases. Current approaches either fit a logistic regression independently to each voxel ignoring any form of spatial dependence or are Bayesian accounting for shared information between neighbouring voxels. However, Bayesian spatial models rely on computationally intensive Markov Chain Monte Carlo (MCMC) methods for inference which are not feasible for large-scale studies. We propose a scalable hierarchical Bayesian spatial model capable of handling binary responses by placing continuous spike-and-slab mixture priors on spatially-varying parameters and enforcing spatial dependency on the parameter dictating the amount of sparsity within the probability of inclusion/exclusion. We use Bayesian Bootstrap for inference accompanied by stochastic variational inference that allows our method to scale to large sample sizes. Moreover, we identify promising sparse high-probability subsets by performing dynamic posterior exploration for structured spike-and-slab regression through approximation of the posterior marginal of latent active variables. Lastly, we validate our results via simulation studies and an application to the UK Biobank.

CO284 Room K0.19 (Hybrid 04) ECONOMETRICS FOR SPORT DATA MODELLING AND FORECASTING Chair: Luca De Angelis

C0463: **Betting on a buzz, mispricing and inefficiency in online sportsbooks**

Presenter: **Carl Singleton**, University of Reading, United Kingdom

Co-authors: James Reade

Bookmakers sell claims to bettors that depend on the outcomes of professional sports events. Like other financial assets, the wisdom of crowds could help sellers to price these claims more efficiently. We use the Wikipedia profile page views of professional tennis players involved in over ten thousand singles matches to construct a buzz factor. This measures the difference between players' pre-match page views relative to the usual number of views they received over the previous year. The buzz factor significantly predicts mispricing by bookmakers. Using this fact to forecast match outcomes, we demonstrate that a strategy of betting on players who received more pre-match buzz than their opponents can generate substantial profits. These results imply that sportsbooks could price outcomes more efficiently by listening to the buzz.

C0601: **In-play betting: Are markets efficient and how do bookmakers make profit?**
Presenter: **Marius Oetting**, TU Clausthal, Germany

High-resolution (1 Hz) data on betting odds and volumes in a football live betting market are considered. We use this unique data set provided by a large European bookmaker to conduct an exploratory analysis of market inefficiencies and the bookmaker's price-setting strategies. Since previous research suggests that mispricing is present after events such as goals, we focus on the pricing of bookmakers and stakes placed by bettors when the first goal of a match is scored. Our results indicate that bettors suffer from well-known behavioural biases, which in turn are exploited by bookmakers in their price setting and thus lead to profit.

C1252: **Underestimating randomness: Outcome bias in betting exchange markets**
Presenter: **Raphael Flepp**, University of Zurich, Switzerland
Co-authors: Oliver Merz, Egon Franck

The purpose is to examine whether the outcome bias harms price efficiency in betting exchange markets. In soccer, the match outcome is an unreliable performance measure, as it underestimates the high level of randomness involved in the sport. If bettors overestimate the importance of past match outcomes and underestimate the influence of good or bad luck, we expect less accurate prices for lucky and unlucky teams. Analyzing over 8,900 soccer matches, we find evidence that the prices are overstated for previously lucky teams and understated for previously unlucky teams. Consistent with the outcome bias, the betting community overestimates the importance of past match outcomes. Consequently, this bias translates into significantly negative betting returns on lucky teams and positive betting returns on unlucky teams. Based on this finding, we propose a simple betting strategy that generates positive returns in an out-of-sample backtest.

C0657: **Home advantage and mispricing in indoor sports' ghost games: The case of European basketball**
Presenter: **Luca De Angelis**, University of Bologna, Italy
Co-authors: James Reade

Several recent studies suggest that the home advantage was - at least temporarily - reduced in ghost games due to the COVID-19 outbreak. However, the majority of these works focus on football and no contributions have been provided for indoor sports, where the effect of the support of the fans might have a stronger impact than in big stadia. We try to fill this gap by investigating the effect of ghost games in basketball with a special focus on the possible reduction of the home advantage due to the absence of spectators inside the arena. In particular, we test (i) for the reduction of the home advantage in basketball, (ii) whether such reduction tends to disappear over time, possibly due to adaptation of players/referees/coaching staff to the new environment, (iii) if the bookmakers promptly adapt to such structural change or whether mispricing was created on the betting market. The results from a large data set covering all seasons since 2004 for the ten most popular and followed basketball leagues in Europe show, on the one hand, an overall significant reduction of the home advantage of around 5% and no evidence that suggests that this effect has softened even after more than 20 rounds; on the other hand, bookmakers appear to have anticipated such effect and priced home win in basketball matches accordingly, thus avoiding creating mispricing on betting markets.

CO394 Room K0.20 (Hybrid 05) ADVANCES IN FINANCIAL NETWORK MODELLING Chair: Cristina Amado

C0420: **Measuring and hedging GEOVOL**
Presenter: **Susana Campos Martins**, University of Oxford, United Kingdom

Some events impact volatilities of most assets, asset classes, sectors and countries, causing serious damage to investment portfolios. The magnitude of such shocks is defined as GEOVOL which is a broad measure of geopolitical risk. The purpose is to introduce a statistical formulation of such events as common volatility innovations in both a multivariate volatility and an asset pricing context. Simulations verify the statistical performance of the simple but novel estimator and a test to detect GEOVOL. Two empirical examples show the events that have had the biggest impact on financial markets. The results are useful for portfolio optimization and risk forecasting.

C1125: **Residual-based nodewise regression in factor models with ultra-high dimensions**
Presenter: **Marcelo Medeiros**, PUC-Rio, Brazil
Co-authors: Mehmet Caner

A new theory is provided for nodewise regression when the residuals from a fitted factor model are used to apply our results to the analysis of the maximum Sharpe ratio when the number of assets in a portfolio is larger than its time span. We introduce a new hybrid model where factor models are combined with feasible nodewise regression. Returns are generated from an increasing number of factors plus idiosyncratic components (errors). The precision matrix of the idiosyncratic terms is assumed to be sparse, but the respective covariance matrix can be non-sparse. Since the nodewise regression is not feasible due to the unknown nature of errors, we provide a feasible-residual-based nodewise regression to estimate the precision matrix of errors as a new method. Next, we show that the residual-based nodewise regression provides a consistent estimate for the precision matrix of errors. In another new development, we also show that the precision matrix of returns can be estimated consistently, even with an increasing number of factors.

C1262: **Global volatility shocks and the PPP puzzle**
Presenter: **Tales Padilha**, University of Oxford, United Kingdom
Co-authors: Susana Campos Martins

Most of the discussion about the Purchasing Power Parity (PPP) Puzzle has pertained to the reversion speed of deviations from PPP. Much less attention, however, has been given to the other component of the puzzle: the high volatilities of real exchange rates. We provide a framework that is capable of explaining the econometric sources of these volatilities. First, we study the drivers of real exchange rate volatilities using a Cross-Sectionally Augmented Autoregressive Distributed Lag (CS-ARDL) panel framework and the conditional covariance matrices of the system with nominal exchange rates and price differentials. This analysis indicates that, for both emerging and developed markets, common factors are the main drivers of volatility. With this result in hand, we propose a novel econometric framework - based on the endogenous common volatility shocks model - that explains the sources of these volatilities as common second-moment shocks. This framework allows us to give structure to the origins of these high volatilities and propose an extension to study their macro-financial drivers.

C1347: **Common asset holdings and systemic vulnerability across multiple types of financial institution**
Presenter: **Paolo Barucca**, University College London, United Kingdom
Co-authors: Laura Silvestri, Tahir Mahmood

One way systemic risk can crystallise is through fire sales of commonly held assets. Fire sale vulnerabilities across different types of financial institutions, including non-banks, are examined. We undertake an in-depth empirical analysis of the interconnections between European open-

ended investment funds and the UK regulated banks and insurance companies through their common asset holdings. The aim is first to combine regulatory holding-level asset data for banks and insurers with private data for open-ended investment funds. Our results show the existence of a significant overlap between the equity and debt portfolios of different types of financial institutions. We characterise financial institutions of different types in terms of their concentration profile, portfolio similarity and vulnerability to fire sales, providing evidence for the existence of a price-mediated channel of contagion between banks, insurance companies and investments funds.

CO386 Room K0.50 (Hybrid 06) MIXED FREQUENCY AND ASSET ALLOCATION Chair: Ekaterina Kazak

C1039: **Testing for endogeneity of irregular sampling schemes**
Presenter: **Aleksey Kolokolov**, Manchester Business School, United Kingdom
Co-authors: Davide Pirino, Giulia Livieri
In the context of high-frequency data, a simplifying assumption of the independence between the sampling scheme and the observed process itself is often made. In order to justify or reject the assumption empirically, we propose a statistical test for the endogeneity of the sampling times. The test is robust to the presence of jumps and can be used for detecting the dependence between zeros in the financial prices sampled at a moderate frequency and the efficient price process. Extensive Monte Carlo simulations confirm the good finite sample performance of the proposed test.

C0326: **In-sample inference for MIDAS regressions and GARCH-MIDAS**
Presenter: **Onno Kleen**, Erasmus University Rotterdam, Netherlands
Co-authors: Andrea Naghi, Michel van der Wel
Identification issues in mixed-frequency data sampling (MIDAS) models are examined. In MIDAS models, data sampled at two different frequencies are typically linked via one single parameter. This parameter is of particular interest in determining whether there is a significant relationship between the MIDAS component and the dependent variable. However, due to nuisance parameters, the distribution of the linking parameter is nonstandard. We discuss possible ways of adjusting the test statistics in unidentified and weakly-identified cases. We apply these robust inference procedures in a simulation study of linear MIDAS models and in empirical re-examinations of previous GARCH-MIDAS applications.

C0913: **Market response to a fear impulse**
Presenter: **Stefan Voigt**, University of Copenhagen, Denmark
Co-authors: Nikolaus Hautsch, Albert Menkveld
Thirty billion Nasdaq order-book messages for exchange-traded funds are analyzed to delineate how the market responds to a VIX impulse. We find that investors actively sell equities and buy government bonds on largely unchanged liquidity. A deeper analysis shows that this result is entirely driven by investors becoming more averse to uncertainty. In other words, changes in the variance-risk premium are responsible for the pattern we find for VIX impulses. For impulses driven by changes in cash-flow risk, we find active buying of equities on worse liquidity. We rationalize these patterns by essentially adding risk shocks to a previous study.

C0942: **Portfolio allocation using echo state networks**
Presenter: **Michael Grebe**, The University of Manchester, United Kingdom
Co-authors: Ekaterina Kazak
Portfolio optimization has been extensively investigated since the 1990s and found widespread application in finance and economics. The state-of-the-art Echo State Network is applied to forecast portfolio weights and improve portfolio allocation. Echo State Networks present a novel approach to the estimation method of Recurrent Neural Networks, reducing computational effort and overcoming commonly faced challenges of Recurrent Neural Networks. The portfolio optimization problem is studied for different portfolio sizes using a dataset on the S&P100 index from the beginning of 2014 to the end of 2019. A dynamic hyperparameter optimization approach is employed and the empirical results show that the Echo State Network outperformed the commonly used weight estimation approaches based on dynamic conditional variance models.

CO581 Room Virtual R20 TOPICS IN TIME SERIES AND FINANCIAL ECONOMETRICS Chair: Alessandra Amendola

C0292: **Parallel computations for nonparametric estimation of risk-neutral densities through option prices**
Presenter: **Antonio Santos**, University of Coimbra, Portugal
Co-authors: Ana Monteiro
The risk-neutral density is one of the key objects in option contracts pricing. One of the most flexible ways to estimate that object, especially within big data environments through intraday data available nowadays, is using nonparametric estimation methods. In this context, two main challenges need to be addressed. First, implementing estimation procedures through large-scale constrained convex optimization problems, more robust when compared with simpler estimators, benchmarked by the Nadaraya-Watson estimator. Second, there is the computational challenge associated with nonparametric estimators for choosing the model's complexity done by defining an optimal bandwidth. The state-of-the-art method for choosing the optimal bandwidth is Cross-Validation, a problem that can be parallelized. The paradigm to implement such procedures is using Graphics Processing Units computational capabilities. We demonstrate the use of such capabilities for defining optimal bandwidths in the nonparametric estimation of risk-neutral densities, where thousands of convex optimization problems are solved in parallel. The application of these computational methods is considered in the estimation of risk-neutral densities associated with option contracts for S&P500 and VIX indexes.

C1635: **Measuring model risk for market risk models**
Presenter: **Emese Lazar**, University of Reading, United Kingdom
Co-authors: Radu Tunaru, Ning Zhang
A scoring function-based model risk estimation methodology is proposed for measuring the joint model risk of Value-at-Risk and Expected Shortfall. We apply the proposed model risk measure across various market risk models and in our simulation study we find that our technique captures a large proportion of true model risk. We show that model risk is not always subadditive and when model risk is present, the ranking of market risk models is affected by the scoring function.

C0729: **VaR modeling: Conditional quantile dependence approaches**
Presenter: **Giorgia Rivieccio**, Parthenope University, Italy
Co-authors: stefania Corsaro, Giovanni De Luca, Javier Ojea Ferreiro
Risk Management typically focuses on the Value-at-Risk (VaR) as the main risk measure. VaR is financially interpreted as the worst expected loss of a portfolio over a specified holding period at a given confidence level (generally 1% or 5%) over one day. Statistically, the estimate of the VaR corresponds to the estimate of a tail quantile of the conditional distribution of future portfolio returns. Its measurement is a highly challenging statistical problem. The classical approaches have been partly overcome by the Multivariate Conditional AutoRegressive specification for VaR (MCAViaR) which directly estimates the dynamics of the quantiles without modeling the distribution of returns. Such an approach considers possible spillovers on the VaRs, assuming a linear model for bivariate conditional quantiles. However, the assumption of linearity is also its limit. We then propose alternative copula-based approaches, specifying both a static and a dynamic dependence structure. To measure the performance in estimating the VaR of a portfolio of financial returns, we have compared three models: the MCAViaR model, a copula-VaR model and a time-varying regime switching copula model.

C1361: **The predictive content of the textual political polarity index: The case of Italian GDP**
Presenter: **Alessandro Grimaldi**, University of Salerno, Italy
Co-authors: Alessandra Amendola, Walter Distaso

A data-driven approach is proposed to derive a Textual Political Polarity Index (*TPPI*) based on the analysis of the entire collection of the verbatim reports of the Italian "*Senate of the Republic*". The procedure allows us to build a set of polarity indices reflecting the impact of political debate - as well as agreement/disagreement within parties' groups - on a specific economic variable over time. In order to assess such an impact, we perform predictive regressions on a chosen macroeconomic variable - namely, the yearly Italian GDP growth rate. Results point to a nontrivial predictive power of the proposed polarity indices, which (importantly) do not rely on a subjective choice of an affective lexicon.

CO720 Room Virtual R31 ADVANCES IN BAYESIAN ECONOMETRICS Chair: Helga Wagner

E0185: **Ultimate Polya gamma samplers: Efficient MCMC for possibly imbalanced binary and categorical data**
Presenter: **Gregor Zens**, Vienna University of Economics and Business, Austria
Co-authors: Sylvia Fruehwirth-Schnatter, Helga Wagner

Modeling binary and categorical data is one of the most commonly encountered tasks of applied statisticians and econometricians. While Bayesian methods in this context have been available for decades now, they often require a high level of familiarity with Bayesian statistics or suffer from issues such as low sampling efficiency. To contribute to the accessibility of Bayesian models for binary and categorical data, we introduce novel latent variable representations based on Polya Gamma random variables for a range of commonly encountered discrete choice models. New Gibbs sampling algorithms for binary, binomial and multinomial logistic regression models are derived from these latent variable representations. All models allow for a conditionally Gaussian likelihood representation, rendering extensions to more complex modeling frameworks such as state-space models straight-forward. However, sampling efficiency may still be an issue in these data augmentation based estimation frameworks. To counteract this, MCMC boosting strategies are developed and discussed in detail. The merits of our approach are illustrated through extensive simulations and a real data application.

E0281: **Generalized mixtures of finite mixtures**
Presenter: **Gertraud Malsiner-Walli**, WU Vienna University of Economics and Business, Austria
Co-authors: Sylvia Fruehwirth-Schnatter, Bettina Gruen

Within a Bayesian framework, an investigation of the model class of mixtures of finite mixtures (MFMs) where a prior on the number of components is specified is performed. This model class requires suitable prior specifications and inference methods to exploit its full potential. We contribute to the Bayesian analysis of MFMs by considering a generalized class of MFMs containing static and dynamic MFMs where the Dirichlet parameter of the component weights either is fixed or depends on the number of components. We emphasize the distinction between the number of components K of a mixture and the number of clusters $K+$, i.e., the number of filled components. In the MFM model, $K+$ is a random variable, and its prior depends on the prior on the number of components K and the mixture weights. We characterize the prior on the number of clusters $K+$ and derive computationally feasible formulas to calculate this implicit prior. For posterior inference, we propose the telescoping sampler, which allows Bayesian inference for mixtures with arbitrary component distributions. The telescoping sampler explicitly samples the number of components, but otherwise requires only the usual MCMC steps for estimating a finite mixture model. The ease of its application is demonstrated on a real data set.

C0282: **Fast and accurate variational inference for large Bayesian VARs with stochastic volatility**
Presenter: **Xuewen Yu**, Purdue University, United States
Co-authors: Joshua Chan

A new variational approximation is proposed for the joint posterior distribution of the log-volatility in the context of large Bayesian VARs. In contrast to existing approaches based on local approximations, the new proposal provides a global approximation that considers the entire support of the joint distribution. A Monte Carlo study shows that the new global approximation is over an order of magnitude more accurate than existing alternatives. We illustrate the proposed methodology with an application of a 96-variable VAR with stochastic volatility to measure global bank network connectedness. Our measure is able to detect the drastic increase in global bank network connectedness much earlier than rolling-window estimates from a homoscedastic VAR.

C0600: **Bayesian optimization of hyperparameters when the marginal likelihood is estimated by MCMC**
Presenter: **Mattias Villani**, Stockholm University, Sweden
Co-authors: Oskar Gustafsson, Par Stockhammar

Bayesian models in econometrics often involve a small set of hyperparameters determined by maximizing the marginal likelihood. Bayesian optimization is a popular iterative method where a Gaussian process posterior of the underlying function is sequentially updated by new function evaluations. An acquisition strategy uses this posterior distribution to decide where to place the next function evaluation. We propose a novel Bayesian optimization framework for situations where the user controls the computational effort, and therefore the precision of the function evaluations. This is a common situation in econometrics where the marginal likelihood is often computed by Markov Chain Monte Carlo (MCMC) methods, with the precision determined by the number of MCMC draws. The method is used to find optimal prior hyperparameters in the steady-state vector autoregression fitted to US macroeconomic data.

CO268 Room Virtual R33 EMPIRICAL MACRO Chair: Laura Jackson Young

C0375: **The role of corporate tax policy on monetary effectiveness: A quasi-experimental approach**
Presenter: **Ezgi Kurt**, Bentley University, United States

The first empirical evidence on how corporate tax policy affects monetary policy outcomes is documented. Using exogenous marginal tax reforms in the US, we show that the average impact of monetary policy differs based on the tax treatments firms receive. Specifically, we find that monetary policy is more effective on employment and investment for firms facing tax increases relative to those with stable statutory taxes. Moreover, we document that monetary policy is least effective when firms face marginal tax cuts. The empirical findings are rationalized using a New Keynesian model featuring capital and corporate income taxes. Both theoretical model and empirical findings suggest that tax policy could lead to considerable variation in monetary policy outcomes.

C0612: **Securitization and bank risk: Evidence from CLO-funded syndicated loans**
Presenter: **Andrea Civelli**, University of Arkansas, United States
Co-authors: Santiago Barraza

The focus is on the causal effect on bank risk of bank reliance on collateralized loan obligations to fund business loan originations. We find that exogenous increases in liquidity in the CLO market significantly decrease bank expected default frequency for half a year, with a peak response about a quarter after the shock. The results highlight the soothing effect of securitization under favorable market conditions, as well as its distressing effect under market duress. The pro-cyclical nature of these effects calls for attention from both bank managers and policy-makers to the ever-growing use of CLOs.

C1106: **A time-varying threshold STAR model of unemployment and the natural rate**
Presenter: **Laura Jackson Young**, Bentley University, United States

Smooth-transition autoregressive (STAR) models, competitors of Markov-switching models, are limited by an assumed time-invariant threshold level. However, a STAR framework could estimate a time-varying threshold level of unemployment. One can consider this threshold a "tipping level" where the mean and dynamics of the natural rate of unemployment shift. If the threshold level is time-varying, one can add an error-correction term—between the lagged levels of unemployment and the threshold — to the autoregressive terms in the STAR model. Thus, the time-varying latent threshold level serves as both a demarcation between regimes and an error-correction term.

C0772: **Comparing monetary policy tools in an estimated DSGE model with International financial markets**
Presenter: **Sacha Gelfer**, Bentley University, United States
Co-authors: Christopher Gibbs

The dynamics of conventional and unconventional monetary policy are evaluated using an estimated a two-region dynamic stochastic general equilibrium (DSGE) model. In addition to traditional nominal frictions, the open-economy model also includes financial frictions, international portfolio balance effects, and correlated global financial shocks. We find that both conventional and unconventional monetary policy is effective in stimulating output and inflation. However, the type of expansionary monetary policy used has heterogeneous effects on domestic investment, imports, exports hours worked and financial markets. Further, including a financial accelerator to the DSGE model significantly dampens the impact of aggregate investment that is expected to occur with unconventional monetary policy. This is because unconventional monetary policy in the model is associated with an expansion in banking deposits and a minimal impact on loan demand, thus creating a fall in the loan to deposit ratio as was seen over the decade after the global financial crisis. Finally, using historical decompositions, we find that global unconventional monetary policy had a significant positive impact on output, exports, and hours worked during the global financial crisis and the preceding years after, but becomes negligible after 2014. Yet, its impact on asset markets and bond markets remained through 2019.

CO679 Room Virtual R34 FORECASTING UNDER STRUCTURAL CHANGE Chair: Andrew Martinez

C1332: **Evaluating the federal reserves tealbook forecasts**
Presenter: **Neil Ericsson**, Federal Reserve Board, United States

Publicly available Federal Reserve Board Tealbook forecasts of GDP growth for the United States and several foreign countries are examined focusing on potential time-varying biases and evaluating the Tealbook forecasts relative to other institutions forecasts. Tealbook forecasts perform relatively well at short horizons, but with significant heterogeneity across countries. Also, while standard Mincer-Zarnowitz tests typically fail to detect biases in the Tealbook forecasts, recently developed indicator saturation techniques that employ machine learning are able to detect economically sizable and highly significant time-varying biases. Estimated biases differ not only over time, but by country and across the forecast horizon. These biases point to directions for forecast improvement. Previous forecast-encompassing tests of the Tealbook forecasts relative to JP Morgan's forecasts reveal distinct value added by each institution's forecasts. However, for most countries and forecast horizons examined, each institution's forecast can be improved by utilizing information from the other institution's forecast.

C1275: **Nonlinear dynamic factor models**
Presenter: **Molin Zhong**, Federal Reserve Board, United States
Co-authors: Pablo Guerron, Alexey Khazanov

A new dynamic factor model is proposed that allows nonlinear dynamics in the state and measurement equations. The proposed nonlinear factor model 1) can generate asymmetric, state-dependent, and size-dependent responses of observables to shocks; 2) can produce time-varying volatility, skewness, and tail risks in the predictive distributions; and 3) fits the data better than a linear factor model. Using macroeconomic and financial variables, we show how to take the model to the data. We find overwhelming evidence in favor of the nonlinear factor model over its linear counterpart in applications that include interest rates with zero lower bounds, credit default swap spreads for European countries, and nonfinancial corporate credit default swap spreads in the U.S.

C1330: **Smooth robust multi-horizon forecasts**
Presenter: **Andrew Martinez**, US Department of the Treasury, United States
Co-authors: Jennifer L Castle, David Hendry

The purpose is to investigate whether smooth robust methods for forecasting can help mitigate pronounced and persistent failure across multiple forecast horizons. We demonstrate that naive predictors are interpretable as local estimators of the long-run relationship with the advantage of adapting quickly after a break, but at a cost of additional forecast error variance. Smoothing over naive estimates helps retain these advantages while reducing the costs, especially for longer forecast horizons. We derive the performance of these predictors after a location shift, and confirm the results using simulations. We apply smooth methods to forecasts of UK productivity and US 10-year Treasury yields and show that they can dramatically reduce persistent forecast failure exhibited by forecasts from macroeconomic models and professional forecasters.

C1344: **Economic forecasting in a shifting world**
Presenter: **David Hendry**, University of Oxford, United Kingdom
Co-authors: Jennifer L Castle, Jurgen Doornik

Economic time series are subject to non-stationarities from both evolving stochastic trends and sudden distributional shifts, often unanticipated such as pandemics. Modelling and forecasting difficulties are exacerbated by the latency in data provision, usually followed by substantive revisions and occasional changes to data measurement systems. Despite a large body of theory, economists have imperfect and incomplete knowledge of their data generating processes from changing human behaviour, so must search for reasonable empirical modelling approximations. Despite such problems, forecasts of likely future outcomes and their uncertainties are essential to plan and adapt as events unfold, over varying horizons for different decisions. We consider how these features shape the formulation and selection of econometric models for forecasting, and apply our tools to forecasting top-income shares.

CO744 Room Virtual R35 EXPECTATIONS AND UNCERTAINTY Chair: Tomasz Lyziak

C0280: **Do expectations drive expectations: On the formation of consumer inflation expectations in the US**
Presenter: **Tomasz Lyziak**, Polish Academy of Science, Poland
Co-authors: Malgorzata Kalbarczyk, Joanna Mackiewicz-Lyziak

Using a novel and unique dataset, the aim is to examine whether consumer inflation expectations in the US depend on expectations regarding other macroeconomic variables. Our results suggest that in order to understand consumer views on future inflation, it is necessary to consider their predictions concerning developments of other macroeconomic variables. Interestingly, expectations regarding fiscal policy appear relevant drivers of consumer inflation expectations, which, jointly with the confidence factor already suggested in the literature, explains the increase of those expectations after the beginning of the Covid-19 pandemics. We also demonstrate that consumers' self-assessments concerning the degree of risk-aversion and health explain differences in inflation expectations among consumers. After considering their subjective expectations and self-assessments, consumers appear to a large extent forward-looking, even if predictability of future inflation developments, as measured with

official statistics, is limited. Also, the dependence of their inflation expectations on selected socio-demographic features, such as income, gender or age, is diminished.

C0569: **Medium- vs. short-term consumer inflation expectations: Evidence from a new euro area survey**
Presenter: **Maritta Paloviita**, Bank of Finland, Finland
Co-authors: Ewa Stanislawska

Using the ECB Consumer Expectations Survey, the purpose is to investigate how consumers revise medium-term inflation expectations. We provide robust evidence of their adjustment to the current economic developments. In particular, consumers adjust medium-term inflation views in response to changes in short-term inflation expectations and, to a lesser degree, to changes in perceptions of current inflation. We find that the strong adverse Covid-19 pandemic shock contributed to an increase in consumer inflation expectations. We show that consumers who declare high trust in the ECB adjust their medium-term inflation expectations to a lesser degree than consumers with low trust. Our results increase understanding of expectations formation, which is an important issue for medium-term oriented monetary policy.

C0838: **Unit cost expectations and uncertainty: Firms' perspectives on inflation**
Presenter: **Xuguang Simon Sheng**, American University, United States
Co-authors: Brent Meyer, Nicholas Parker

Relying on the Atlanta Fed's Business Inflation Expectations Survey, inference about firm's inflation perceptions, expectations, and uncertainty through the lens of firms' unit (marginal) costs is drawn. Using methods grounded in the survey literature, we find evidence that the concept of "aggregate inflation" as measured through price statistics like the Consumer Price Index (CPI) hold very little relevance for business decision-makers. This lack of relevance manifests itself through experiments (including randomized controlled trials) that show varying question-wording researchers use to elicit inflation expectations and perceptions significantly changes firm's responses. The results suggest firms have become rationally ignorant of the concept of inflation in a low inflation environment. Instead, we find that unit (marginal) costs are the relevant lens with which to capture firms' views on the nominal size of the economy.

C1134: **Do the ECBs introductory statements help predict monetarypolicy? Evidence from a tone analysis**
Presenter: **Pawel Baranowski**, University of Lodz, Poland
Co-authors: Hamza Bennani, Wirginia Doryn

The aim is to examine whether a tone shock derived from European Central Bank communication helps predict ECB monetary policy decisions. To this purpose, we first use a bag-of-words approach and several dictionaries on the ECBs Introductory Statements to derive a measure of tone. Next, we orthogonalise the tone measure on the latest data available to market participants to compute the tone shock. Finally, we relate the tone shock to future ECB monetary policy decisions. We find that the tone shock is significantly and positively related to future ECB monetary policy decisions, even when controlling for market expectations of monetary policy and the Governing Councils inter-meeting communication. Further extensions show that the predictive ability of the tone shock is robust to (i) the normalization of the tone measure, (ii) alternative market expectations of monetary policy, and (iii) the horizon of macroeconomic variables used in the Taylor-type monetary policy rule. These findings highlight an additional channel through which ECB communication improves monetary policy predictability, suggesting that the ECB may have private information that it communicates through its Introductory Statements.

CO511 Room Virtual R36 HIGH-DIMENSIONAL PORTFOLIO SELECTION Chair: Yarema Okhrin

C0566: **Nonlinear interconnectedness of crude oil and financial markets**
Presenter: **Yarema Okhrin**, University of Augsburg, Germany
Co-authors: Gazi Salah Uddin

The purpose is to investigate the heterogeneous and asymmetrical effect of COVID-19 on the crude oil, SP500 index, EUR/USD exchange rate, and various uncertainty measures. These assets reflect the overall health of the global financial and economic system. The COVID-19 pandemic has contributed significantly to demand and supply shocks that have led to an unprecedented decline in crude oil prices. The heterogeneous and asymmetric impact of COVID-19 on these different asset classes is examined. This would enable us to understand how different asset classes react to such unique shocks. The contribution is fourfold. First, we evaluated the impact of the COVID-19 crisis on the interconnectedness of the financials, forex, and commodity markets with a specific focus on risk dynamics. Second, in contrast to the previous studies we consider high-frequency intraday data. This allows us to provide a deeper insight into the dependencies at a daily level. Third, we quantify the dependence and its dynamics using paired vine copulas. This class of copulas is highly flexible and can allow for a convenient visualization of the dependence. Forth, we put a particular focus on the crude oil returns as a function of several financial covariates using C- and D-vine regressions. This approach allows us to model the whole conditional distribution within a single day and to get insights into the causal dependence in tails or at particular quantiles.

C0448: **Dynamic shrinkage estimation of the high-dimensional minimum-variance portfolio**
Presenter: **Erik Thorsen**, Stockholm University, Sweden
Co-authors: Taras Bodnar, Nestor Parolya

New results in random matrix theory are derived which allow the construction of a shrinkage estimator of the global minimum variance (GMV) portfolio when the shrinkage target is a random object. More specifically, the shrinkage target is determined as the holding portfolio estimated from previous data. The theoretical findings are applied to develop theory for dynamic estimation of the GMV portfolio, where the new estimator of its weights is shrunk to the holding portfolio at each time of reconstruction. Both cases with and without overlapping samples are considered. The non-overlapping samples correspond to the case when different data of the asset returns are used to construct the traditional estimator of the GMV portfolio weights and to determine the target portfolio, while the overlapping case allows intersections between the samples. The theoretical results are derived under weak assumptions imposed on the data-generating process. No specific distribution is assumed for the asset returns except from the assumption of finite $4+\varepsilon$, $\varepsilon > 0$, moments. Also, the population covariance matrix with unbounded spectrum can be considered. The performance of new trading strategies is investigated via an extensive simulation. Finally, the theoretical findings are implemented in an empirical illustration based on the returns on stocks included in the S&P 500 index.

C0632: **Is the empirical out-of-sample variance an informative risk measure for the high-dimensional portfolios?**
Presenter: **Nestor Parolya**, Delft University of Technology, Netherlands
Co-authors: Taras Bodnar, Erik Thorsen

The main contribution is the derivation of the asymptotic behaviour of the out-of-sample variance, the out-of-sample relative loss, and of their empirical counterparts in the high-dimensional setting, i.e., when both ratios p/n and p/m tend to some positive constants as m and n approach infinity, where p is the portfolio dimension, n and m are the sample sizes from the in-sample and out-of-sample periods, respectively. The results are obtained for the traditional estimator of the GMV portfolio, for two previous shrinkage estimators, and for the equally-weighted portfolio, which is used as a target portfolio in the specification of the two considered shrinkage estimators. We show that the behaviour of the empirical out-of-sample variance may be misleading in many practical situations. On the other hand, this will never happen with the empirical out-of-sample relative loss, which seems to provide a natural normalization of the out-of-sample variance in the high-dimensional set-up. As a result, an important question arises if this risk measure can be safely used in practice for portfolios constructed from a large asset universe.

C0919: **Volatility sensitive Bayesian estimation of portfolio VaR and CVaR**
Presenter: **Vilhelm Niklasson**, Stockholm University, Sweden
Co-authors: Taras Bodnar, Erik Thorsen

A new way to integrate volatility information for estimating value at risk (VaR) and conditional value at risk (CVaR) of a portfolio is suggested. The new method is developed from the perspective of Bayesian statistics and it is based on the idea of volatility clustering. By specifying the hyperparameters in a conjugate prior based on two different rolling window sizes, it is possible to quickly adapt to changes in volatility and automatically specify the degree of certainty in the prior. This constitutes an advantage in comparison to existing Bayesian methods that are less sensitive to such changes in the market and also usually lack standardized ways of expressing the degree of belief. We illustrate our new approach using both simulated and empirical data and conclude that it provides a good alternative for risk estimation, especially during turbulent periods.

CO300 Room K2.41 (Hybrid 09) TEXT MINING AND SENTIMENT ANALYSIS FOR ECONOMICS AND FINANCE Chair: Manuela Pedio

C0774: **Sentiment analysis of economic text: A lexicon-based approach**
Presenter: **Elisa Tosetti**, University of Venice, Italy
Co-authors: Luca Tiozzo Pezzoli, Luca Barbaglia, Sergio Consoli, Sebastiano Manzan

With the increasing availability of opinion-rich web resources, such as news, discussion forum and personal blogs, a growing body of research in economics and finance focuses on constructing sentiment indicators from these sources with the aim to predict in a timely manner economic and financial developments. Several studies automatically determine the sentiment of a piece of text by looking at how many positive and negative words can be found in the text according to a predefined lexicon. While this approach is very popular, little work has been done to develop a lexicon for calculating sentiment scores specifically for text with economic content. We fill this gap by proposing a domain-specific lexicon suitable for applications in the area of economics and finance. We do this by carrying a semantic analysis of the language used in economic news and extracting the set of most frequent terms used to describe economic concepts. We use the proposed lexicon in a small empirical study to investigate to what extent the forecasting power of a regression model for predicting stock returns can be improved by incorporating sentiment extracted from economic text.

C0528: **Financial forecasting with word embeddings extracted from news: A preliminary analysis**
Presenter: **Sergio Consoli**, Joint Research Centre (JRC), Italy
Co-authors: Luca Barbaglia

News represents a rich source of information about financial agents actions and expectations. We rely on word embedding methods to summarize the daily content of news. We assess the added value of the word embeddings extracted from US news, as a case study, by using different language approaches while forecasting the US S&P500 index by means of DeepAR, an advanced neural forecasting method based on auto-regressive Recurrent Neural Networks operating in a probabilistic setting. Although this is currently ongoing work, the obtained preliminary results look promising, suggesting an overall validity of the employed methodology.

C0847: **Forecasting realized equity volatility from text sentiment revealed by company filings**
Presenter: **Massimo Guidolin**, Universita' Commerciale Luigi Bocconi - BAFFI CAREFIN, Italy
Co-authors: Manuela Pedio

The aim is to analyze the effect of sentiment on realized equity return volatility in the week following the filing of a 10-K document by the company. We combine the terms in the word lists developed in a previous paper, using a market-based weighting scheme to summarize word frequencies into one sentiment measure. We find that negative, positive, assertive, and litigious tones in the 10-Ks filings have a significant impact on post-filing realized volatility. Our results also show that a market-based weighting scheme produces more reliable results compared to traditional, corpus-based approaches.

C1533: **Mutual fund trust: Evidence from qualitative disclosure**
Presenter: **Liying Wang**, City, University of London, United Kingdom

The level of complexity that mutual funds adopt in their communications to existing and potential investors is studied. We find that low-quality funds manipulate their prospectuses, making them more complex, possibly targeting less sophisticated investors. These investors, in turn, use a less sophisticated asset pricing model to evaluate fund performance, react more aggressively to past winners, and are less sensitive to fund fees. We also find that funds with a relatively low-complexity prospectus are likely to outperform other funds, and their fees are more in line with their managerial skills. The results suggest that funds with low-complexity prospectuses are more trustworthy, and that funds with high-complexity prospectuses are subject to more severe agency issues.

CG586 Room K2.31 Nash (Hybrid 07) CONTRIBUTIONS IN APPLIED MACHINE LEARNING Chair: Klaus Holst

C1671: **Dynamic decision making with reinforcement learning**
Presenter: **Lukas Vacha**, Institute of Information Theory and Automation of the CAS, Czech Republic
Co-authors: Jozef Barunik

A solution is proposed to a general class of models under uncertainty with agents having quantile preferences and limited information processing capacity. We demonstrate our reinforcement learning approach on a simple example where the agent acquires an optimal amount of information. The agent has a limited amount of attention since the information he obtains is costly. Our method can be further extended to more complicated high-dimensional problems, where an analytical solution is impossible to obtain, whereas our reinforcement learning approach makes this task computationally feasible.

C1682: **Evolution of topics in central bank speech communication**
Presenter: **Magnus Hansson**, University of Gothenburg, Sweden

The content of central bank speech communication from 1997 through 2020 is studied and the following questions are asked: (i) What global topics do central banks talk about? (ii) How do these topics evolve over time? We turn to natural language processing, and more specifically Dynamic Topic Models, to answer these questions. The analysis consists of an aggregate study of nine major central banks and a case study of the Federal Reserve, which allows for region-specific control variables. We show that: (i) Central banks address a broad range of topics. (ii) The topics are well captured by Dynamic Topic Models. (iii) The global topics exhibit strong and significant autoregressive properties not easily explained by financial control variables.

C1773: **International cross industry return predictability: Evidence from the US, UK and China**
Presenter: **Yawen Zheng**, University of Liverpool, United Kingdom
Co-authors: Michael Ellington, Michalis Stamatogiannis

The adaptive LASSO and Double LASSO from the statistical learning literature are used to identify economic links between domestic and international industry portfolios. The frameworks allow for complex international industry interdependencies. We find extensive evidence that lagged returns of foreign industries are important when forecasting domestic ones, which is consistent with the gradual diffusion of information hypothesis. We show this using a set of three trading partners, the US, UK and China. In response to the out of sample critique in the stock return predictability literature, we find that utilising a combination forecasting approach with the international information leads to significant out of sample gains.

C1245: **Policy estimation using realistic Q-learning with applications to asset management**

Presenter: **Klaus Holst**, A.P. Moller-Maersk, Denmark

Co-authors: Andreas Nordland

Maintenance of equipment is essential for the profitability of asset-heavy industries in construction, manufacturing and logistics. In general, existing maintenance policies will not be implemented with data-driven optimization in mind. Thus, improvements must rely on the available observational data. This imposes challenges for identifying the optimal policy, which has largely been ignored in management science. A novel application of statistical policy learning for physical asset management is presented, clearly stating the structural assumptions needed to ensure causal interpretability of the findings. Specifically, we estimate an optimal maintenance policy for refrigerated containers owned by the shipping company Maersk. The application requires special consideration for avoiding policy violations. Thus, we apply realistic Q-learning, which adjusts for the limited variability in the decision process induced by the existing maintenance guidelines. The fitted policy's value is validated using double machine learning, and we show a significant gain in value under the fitted optimal policy.

Saturday 18.12.2021 16:30 - 18:35 Parallel Session E – CFE-CMStatistics

EI022 Room K E. Safra (Multi-use 01) BAYESIAN MODEL AND VARIABLE SELECTION (HYBRID) Chair: Francisco Javier Rubio

E0158: **Improper models for data analysis**
Presenter: **David Rossell**, Universitat Pompeu Fabra, Spain
Co-authors: Jack Jewson

Statisticians often face the choice between using probability models or a paradigm defined by minimising a loss function. Both approaches are useful and, if the loss can be re-cast into a proper probability model, there are many tools to decide which model or loss is more appropriate to explain the data's nature. However, when the loss leads to an improper model, there are no principled ways to guide this choice. We address this task by combining the Hyvrinen score, which naturally targets infinitesimal relative probabilities, and general Bayesian updating, which provides a unifying framework for inference on losses and models. Specifically, we propose the H-score, a general Bayesian selection criterion and prove that it consistently selects the (possibly improper) model closest to the data-generating truth in Fisher's divergence. We also prove that an associated H-posterior consistently learns optimal hyper-parameters featuring in loss functions, including a challenging tempering parameter in generalised Bayes / Gibbs posteriors / PAC Bayes. As examples, we consider robust regression and non-parametric density estimation, where popular loss functions define improper models for the data. We hence cannot be dealt with using standard model selection tools. These examples illustrate advantages in robustness-efficiency trade-offs and provide a Bayesian implementation for kernel density estimation, opening a new avenue for Bayesian non-parametrics.

E0159: **Variable selection in mixture models: Uncovering cluster structures and relevant features**
Presenter: **Mahlet Tadesse**, Georgetown University, United States

Identifying latent classes and component-specific relevant predictors can shed important insights when analyzing high-dimensional data. Methods to address this problem in a unified manner will be presented by combining mixture models and variable selection ideas in different contexts. In particular, we will discuss (1) a bi-clustering approach that allows clustering on subsets of variables by introducing latent variable selection indicators in finite or infinite mixture models, (2) an integrative model to relate two high-dimensional datasets by fitting a multivariate mixture of regression models using stochastic partitioning, and (3) a mixture of regression trees approach to uncover homogeneous subgroups and their associated predictors accounting for non-linear relationships and interaction effects. We will illustrate the methods with various genomic applications.

E0160: **Variable selection via Thompson sampling**
Presenter: **Veronika Rockova**, University of Chicago, United States

Thompson sampling is a heuristic algorithm for the multi-armed bandit problem which has a long tradition in machine learning. The algorithm has a Bayesian spirit in the sense that it selects arms based on posterior samples of reward probabilities of each arm. By forging a connection between combinatorial binary bandits and spike-and-slab variable selection, we propose a stochastic optimization approach to subset selection called Thompson Variable Selection (TVS). TVS is a framework for interpretable machine learning which does not rely on the underlying model to be linear. TVS brings together Bayesian reinforcement and machine learning in order to extend the reach of Bayesian subset selection to non-parametric models and large datasets with very many predictors and/or very many observations. Depending on the choice of a reward, TVS can be deployed in offline as well as online setups with streaming data batches. Tailoring multiplay bandits to variable selection, we provide regret bounds without necessarily assuming that the arm mean rewards be unrelated. We show a very strong empirical performance on both simulated and real data. Unlike deterministic optimization methods for spike-and-slab variable selection, the stochastic nature makes TVS less prone to local convergence and thereby more robust.

EO609 Room K0.18 (Hybrid 03) ESTIMATING TREATMENT EFFECTS: METHODS AND APPLICATIONS Chair: David van Dyk

E0809: **A nonparametric doubly robust test for a continuous treatment effect**
Presenter: **Charles Doss**, University of Minnesota, United States
Co-authors: Guangwei Weng, Lan Wang

The vast majority of literature on evaluating the significance of a treatment effect based on observational data has been confined to discrete treatments. These methods are not applicable to drawing inference for a continuous treatment, which arises in many important applications. To adjust for confounders when evaluating a continuous treatment, existing inference methods often rely on discretizing the treatment or using (possibly misspecified) parametric models for the effect curve. To the best of our knowledge, a completely nonparametric doubly robust approach for inference in this setting is not yet available. We develop such a nonparametric doubly robust procedure for making inferences on the continuous treatment effect curve. Using empirical process techniques for local U- and V-processes, we establish the test statistic's asymptotic distribution. Furthermore, we propose a wild bootstrap procedure for implementing the test in practice. We illustrate the new method via simulations and a study of a constructed dataset relating the effect of nurse staffing hours on hospital performance.

E0902: **Disentangling confounding and nonsense associations due to dependence**
Presenter: **Elizabeth Ogburn**, Johns Hopkins University, United States

Nonsense associations can arise when an exposure and an outcome of interest exhibit similar patterns of dependence. Confounding is present when potential outcomes are not independent of treatment. How confusion about these two phenomena underpins popular methods in three areas will be described: causal inference with multiple treatments and unmeasured confounding; causal and statistical inference with social network data; and statistical genetics methods for dealing with unmeasured confounding.

E1056: **Bayesian tree models with targeted smoothing for causal inference**
Presenter: **Jared Murray**, University of Texas at Austin, United States

Bayesian tree models like Bayesian additive regression trees (BART) and Bayesian causal forests (BCF) are popular and effective methods for inferring heterogeneous causal effects. However, their function estimates are necessarily discontinuous and "rough" in their arguments, a significant disadvantage in applications involving continuous treatments or effect moderators thought to have smoothly evolving relationships with treatment efficacy. We extend Bayesian tree models with "targeted smoothing" to allow for (possibly) irregularly spaced continuous treatment variables or moderators while maintaining computational efficiency through the use of carefully constructed basis expansions.

E0620: **Evaluating the impact of built environment interventions in philadelphia**
Presenter: **Shane Jensen**, The Wharton School of the University of Pennsylvania, United States

Urban analytics has recently been improved through publicly available high-resolution data, allowing us to empirically investigate urban design principles of the past half-century. We will evaluate a specifically built environment intervention, the greening of vacant lots in Philadelphia, in terms of potential effects on neighborhood safety and real estate value. We will use several matching strategies to address the issue that greened vacant lots are substantially different from ungreened vacant lots in terms of their surrounding demographic and economic context. We estimate larger and more significant crime reductions around vacant lots that are greened in our matching analysis compared to unmatched analyses. The

effects of vacant lot greening on crime are larger in areas with high residential and low commercial land use and are moderated by the presence of different types of nearby businesses.

E1199: **Online experimentation for studying political polarization**
Presenter: **Alexander Volfovsky**, Duke University, United States

Social media sites are often blamed for exacerbating political polarization by creating echo chambers that prevent people from being exposed to information that contradicts their preexisting beliefs. We conducted a field experiment during which a large group of Democrats and Republicans followed bots that retweeted messages by elected officials and opinion leaders with opposing political views. Republican participants expressed substantially more conservative views after following a liberal Twitter bot, while Democrats attitudes became slightly more liberal after following a conservative Twitter bot although this effect was not statistically significant. As part of a follow up to this experiment, we study the impact of the Russian Internet Research Agency's (IRA) online influence campaign. We find no evidence that interacting with the IRA accounts substantially impacted 6 political attitudes and behaviors. Descriptively, interactions with trolls were most common among individuals who use Twitter frequently, have strong social-media echo chambers, and high interest in politics. We conclude by describing several ongoing field experiments that are designed to elucidate the underlying causes of polarization as well as provide strategies for mitigating it. We will highlight the important causal problems we must solve in order to properly design randomized and observational studies for these complex applied questions.

EO198 Room K0.19 (Hybrid 04) ANALYSIS OF LARGE DATA SETS FOR IMPROVING HEALTHCARE AND TAXATION Chair: Roy Welsch

E0833: **Machine learning methods for survival analysis to predict dementia risk in diabetic patients**
Presenter: **Aamna AlShehhi**, Khalifa University, United Arab Emirates

Dementia is an insidious, progressive, and degenerative neurodegenerative disease that destroys normal brain functionality. It targets the elderly, although it is not part of the normal ageing process. Disease symptoms start with memory loss and language problems that progress over time to losing the ability to carry on normal daily activities. At the later stage, the patient becomes bed-bound and requires around-the-clock care. According to the World Health Organization (WHO), approximately 50 million people worldwide are diagnosed with dementia, with nearly 10 million new cases annually. Dementia has a physical, emotional, financial, and economic burden on the patient as well as on society, families, and caregivers. According to WHO, the estimated global community cost of dementia caring was US$ 818 billion in 2015. On the bright side, dementia can be delayed or prevented by diagnosing it in its early stage. That is why we assessed the performance of different machine learning for survival analysis methods combined with various feature selection methods by the concordance index (C-Index) to predict patients at the risk of developing dementia. In the presented study, we developed a stable predictive model for early-stage dementia prediction using tree-based methods.

E0836: **Challenges and opportunities of competing risks for causal inference**
Presenter: **Bella Vakulenko-Lagun**, University of Haifa, Israel

The motivation comes from research on drug repurposing for Alzheimer's Disease (AD). Any research on AD has to account for competing death that might preclude the onset of AD. We consider a problem of estimation of causal effects of a point intervention from observational data with confounding. While the guidelines for the selection of confounders in a non-competing risks framework are well established, the literature on the selection of confounders in a competing risks setting is extremely sparse - there is only one paper that addresses this problem. We employ specifics of competing risks and explore ways to adjust for confounding. Our results provide the insight and guidelines for practical causal inference applications where competing events cannot be ignored.

E0969: **Do tax deductions encourage charitable giving behavior in the canton of Geneva?**
Presenter: **Marta Pittavino**, University of Geneva, Switzerland
Co-authors: Giedre Lideikyte Huber

Under the current Swiss law, taxpayers can deduct charitable donations from their taxable income (individuals) or profits (corporations) subject to a 20%-threshold. This deductible threshold was increased and introduced in 2006, as part of a larger reform of the Swiss federal tax law, replacing the previous 10%-threshold. The goal of the reform was to boost charitable giving to non-profit entities. However, the efficiency of this reform, and more generally of the existing Swiss system of tax deductions for charitable giving has never been evaluated. Using unique panel data, shared by the Geneva Tax Administration for a time framework of 11 years, an in-depth statistical analysis was conducted. Overall gross income, wealth, together with the year of birth, were the main covariates of interest. Several linear regression models were performed and significant variables, which help answer the questions of taxpayers charitable giving behavior, were identified. Taxpayers were divided into six categories according to the income distribution. We studied the changes in the volume of deductions between categories. The aim is to provide as many taxation insights as possible into both the effects of the 2006 reform, as well as into the patterns of giving and deducting by different classes of taxpayers by income and wealth. The purpose is to provide both Swiss and foreign academics and policymakers with new research and policy insights.

E0976: **Realworld characterizationof blood glucose control and insulinuse in the intensive care unit using EHR data**
Presenter: **Stan Finkelstein**, MIT, United States
Co-authors: Lawrence Baker, Aldo Arevalo, Francis DeMichelle, Roselyn Mateo-Collado, Jason Maley, Leo Celi

The heterogeneity of critical illness complicates both clinical trial design and real-world management. This complexity has resulted in conflicting evidence and opinion regarding optimal management in many intensive care scenarios. Understanding this heterogeneity is essential to tailoring management to individual patients. Hyperglycaemia is one such complication in the intensive care unit (ICU), accompanied by decades of conflicting evidence around management strategies. We hypothesized that analysis of highly-detailed electronic medical record (EMR) data would demonstrate that patients vary widely in their glycaemic response to critical illness and response to insulin therapy. Due to this variability, we believed that hyper- and hypoglycaemia would remain common in ICU care despite standardised approaches to management. We utilized the Medical Information Mart for Intensive CareIII v1.4 (MIMIC) database. We identified 19,694 admissions between 2008 and 2012 with available glucose results and insulin administration data. We demonstrate that hyper- and hypoglycaemia are common at the time of admission and remain so 1 week into an ICU admission. Insulin treatment strategies vary significantly, irrespective of blood glucose level or diabetic status. We reveal a tremendous opportunity for EMR data to guide tailored management. We have made available a highly-detailed data source for future investigation.

E1122: **Is deep reinforcement learning ready for practical applications in healthcare?**
Presenter: **Zach Shahn**, IBM Research, United States
Co-authors: MingYu Lu, Li-wei Lehman

The potential of Reinforcement Learning (RL) has been demonstrated through successful applications to games such as Go and Atari. However, while it is straightforward to evaluate the performance of an RL algorithm in a game setting by simply using it to play the game, evaluation is a major challenge in clinical settings where it could be unsafe to follow RL policies in practice. Thus, understanding sensitivity of RL policies to the host of decisions made during implementation is an important step toward building the type of trust in RL required for eventual clinical uptake. We perform a sensitivity analysis on a state-of-the-art RL algorithm (Dueling Double Deep Q-Networks) applied to hemodynamic stabilization treatment strategies for septic patients in the ICU. We consider the sensitivity of learned policies to input features, embedding model architecture, time discretization, reward function, and random seeds. We find that varying these settings can significantly impact learned policies, which suggests a need for caution when interpreting RL agent output.

EO685 Room K0.20 (Hybrid 05) SCIENTIFICALLY MOTIVATED SPATIAL DATA MODELS Chair: Garritt Page

E0179: **Predicting the risk of novel pathogen introductions from disease surveillance data**
Presenter: **Nelson Walker**, Kansas State University, United States
Co-authors: Trevor Hefley, Daniel Walsh, Ian McGahan, Daniel Skinner, Daniel Storm

In the course of an infectious disease outbreak, researchers often must estimate or infer the source of the causative pathogen, the risk factors associated with the spread and growth of the pathogen, and risk factors that may be associated with new outbreaks. Because the exact time and location of introduction for the pathogen is usually unobserved, these questions must be addressed using incomplete or indirect data, such as spatio-temporal disease surveillance data. We introduce a Bayesian hierarchical mixture model for spatio-temporal, binary disease surveillance data that accounts for the dynamic process of the pathogen diffusing and multiplying through a population from multiple sources. Our framework provides approximate posterior estimates for the number, locations, and times of introduction of the pathogen in a population, as well as posterior inference on parameters associated with pathogen growth and diffusion. We also obtain posterior inference on the generative spatial process that produced the pathogen introductions. We demonstrate this framework using disease surveillance data for chronic wasting disease in white-tailed deer from Wisconsin and Illinois in the USA.

E0285: **Distributional validation of precipitation data products with spatially varying mixture models**
Presenter: **Lynsie Warr**, University of California Irvine, United States
Co-authors: Matthew Heaton, William Christensen, Philip White, Summer Rupper

The high mountain regions of Asia contain more glacial ice than anywhere on the planet outside of the polar regions. Because of the large population living in the Indus watershed region who are reliant on melt from these glaciers for freshwater, understanding the factors that affect glacial melt along with the impacts of climate change on the region is important for managing these natural resources. While there are multiple climate data products (e.g. reanalysis and global climate models) available to study the impact of climate change on this region, each product will have a different amount of skill in projecting a given climate variable, such as precipitation. We develop a spatially varying mixture model to compare the distribution of precipitation in the High Mountain Asia region as produced by climate models with the corresponding distribution from in situ observations from the Asian Precipitation Highly Resolved Observational Data Integration Towards Evaluation (APHRODITE) data product. Parameter estimation is carried out via a computationally efficient Markov chain Monte Carlo algorithm. Each estimated climate distribution from each climate data product is then validated against APHRODITE using a spatially varying Kullback-Leibler divergence measure.

E0549: **Constructing mechanistic spatial models from Ornstein-Uhlenbeck processes**
Presenter: **Nathan Wikle**, University of Texas at Austin, United States
Co-authors: Ephraim Hanks, Corwin Zigler

A mechanistic model is developed to analyze the impact of sulfur dioxide emissions from coal-fired power plants on average sulfate concentrations in the central United States. A multivariate Ornstein-Uhlenbeck (OU) process is used to approximate the dynamics of a linear space-time SPDE. The distributional properties of the OU process are leveraged to specify novel probability models for spatial data (i.e., spatially-referenced data with no temporal replication) that are viewed as either a snapshot or a time-averaged observation of the OU process. Air pollution transport dynamics determine the mean and covariance structure of our atmospheric sulfate model, allowing us to infer which process dynamics are driving observed air pollution concentrations. We use these inferred dynamics to assess the regulatory impact of flue-gas desulfurization (FGD) technologies on human exposure to sulfate aerosols. Extensions of this methodology are discussed, including its potential applicability to methods for causal inference with interference.

E0808: **Detecting changes in dynamic social networks based on unlabeled movement data**
Presenter: **Henry Scharf**, San Diego State University, United States

The social structure of a population can often influence movement and inform researchers on a species' behavioral tendencies. Social networks can be studied through movement data; however, modern sources of data can have complex patterns of missingness that are not straightforward to address using existing methods. For example, drone-gathered observations of trajectories, while highly precise, can introduce labeling issues when individuals in a study population move in and out of the camera's active field of view. When individuals cannot be uniquely identified visually, multiple labels may be assigned to a single individual. Since all available social movement models rely on unique identification of all individuals in the population, we extend an existing Bayesian hierarchical movement model that makes use of a latent social network to accommodate "multiply-labeled" movement data. We apply our model to drone-gathered observations of dolphins to study the effect of sonar exposure on the dolphins social structure. Our proposed framework can be applied to all unlabeled movement data for various social movement applications and has potential implications for the study of privacy-protected movement data.

E1066: **A discretized projection-based method for modeling high-dimensional zero-inflated spatial data**
Presenter: **Seiyon Lee**, George Mason University, United States
Co-authors: Murali Haran

Applications of spatial observations with excessive zeros occur in many disciplines. Modeling such zero-inflated spatial data is computationally challenging, especially in high dimensions. The computational challenge is borne out of inferring the high-dimensional spatial random effects and matrix operations on dense covariance matrices. Markov chain Monte Carlo (MCMC) algorithms may be slow mixing for these models. We propose a computationally efficient approach to model high-dimensional zero-inflated spatial observations using a discretized projections-based approach. Our approach improves mixing in MCMC algorithms and considerably decreases computational overhead for fitting these models. Through simulated examples, we show that our approach performs well in inference and prediction. We also apply our approach to real-world examples in ecology and glaciology.

EO577 Room K0.50 (Hybrid 06) RECENT ADVANCES IN CAUSAL INFERENCE (VIRTUAL) Chair: Andrew Spieker

E0795: **Nonparametric estimation of heterogeneous causal effects**
Presenter: **Edward Kennedy**, Carnegie Mellon University, United States

Heterogeneous effect estimation plays a crucial role in causal inference, with applications across medicine and social science. Many methods for estimating conditional average treatment effects (CATEs) have been proposed in recent years, but there are important theoretical gaps in understanding if and when such methods are optimal. This is especially true when the CATE has a nontrivial structure (e.g., smoothness or sparsity). Two recent papers in this context are surveyed. First, we study a two-stage doubly robust CATE estimator and give a generic model-free error bound, which, despite its generality, yields sharper results than those in the current literature. The second contribution is aimed at understanding the fundamental statistical limits of CATE estimation. To that end, we resolve this long-standing problem by deriving a minimax lower bound, with a matching upper bound based on higher-order influence functions.

E1257: **Semi-parametric estimation of biomarker age trends with endogenous medication use in longitudinal data**
Presenter: **Andrew Spieker**, Vanderbilt University Medical Center, United States
Co-authors: Joseph Delaney, Robyn McClelland

In cohort studies, non-random medication use can pose barriers to the estimation of the natural history trend in a mean biomarker value (namely, the association between a predictor of interest and a biomarker outcome that would be observed in the total absence of biomarker-specific treatment).

Common causes of treatment and outcomes are often unmeasured, obscuring our ability to easily account for medication use with assumptions commonly invoked in causal inference such as conditional ignorability. Without confidence in the availability of a variable satisfying the exclusion restriction, the use of instrumental variable approaches may be difficult to justify. Heckman's hybrid model with structural shift can be used to correct endogeneity bias via a homogeneity assumption and parametric specification of a joint model for the outcome and treatment. The application of this methodology to settings of longitudinal data remains unexplored. We demonstrate how the assumptions of the treatment effects model can be extended to accommodate clustered data arising from longitudinal studies. The proposed approach is semi-parametric in nature in that valid inference can be obtained without the need to specify any component of the longitudinal correlation structure and can serve as a useful tool to uncover natural history trends in longitudinal data that are obscured by endogenous treatment.

E0685: **Evidence factors from multiple, possibly invalid, instrumental variables**
Presenter: **Youjin Lee**, Brown University, United States
Co-authors: Anqi Zhao, Dylan Small, Bikram Karmakar

Instrumental variables have been widely used to estimate the causal effect of a treatment on an outcome in the presence of unmeasured confounders. When several instrumental variables are available and the instruments are subject to possible biases that do not completely overlap, a careful analysis based on these several instruments can produce orthogonal pieces of evidence (i.e., evidence factors) that would strengthen causal conclusions when combined. We develop several strategies, including stratification, to construct evidence factors from multiple candidate instrumental variables when invalid instruments may be present. The proposed methods deliver nearly independent inferential results each from candidate instruments under the more liberally defined exclusion restriction than the previously proposed reinforced design. We apply our stratification method to evaluate the causal effect of malaria on stunting among children in Western Kenya using three nested instruments that are converted from a single ordinal variable. The proposed stratification method is particularly useful when we have an ordinal instrument of which validity depends on different values of the instrument.

E0709: **Bayesian joint modeling for causal mediation analysis with a binary outcome and a binary mediator**
Presenter: **Genevieve Lefebvre**, Universite du Quebec a Montreal, Canada
Co-authors: Miguel Caubet Fernandez, Mariia Samoilenko

Mediation analysis with a binary outcome is notoriously more challenging than with a continuous outcome. We will present a new approach for performing causal mediation with a binary outcome and a binary mediator. Our proposal relies on the Student-t approximation to the Bayesian multivariate regression logistic model. We will explain how this latent multivariate model can be used to estimate the natural direct and indirect effects of an exposure on an outcome in any measuring scale of interest (e.g., odds or risk ratio, risk difference). The novel mediation approach has several valuable features which, to our knowledge, are not found together in current binary-binary mediation approaches. The model will be illustrated and compared to two existing approaches for conducting causal mediation analyses with this type of data.

E1088: **Cluster randomized trials: Assumptions, estimands, and estimation in the presence of post-randomization selection**
Presenter: **Georgia Papadogeorgou**, University of Florida, United States
Co-authors: Fan Li, Fan Li

In cluster-randomized trials, treatment is assigned randomly at the level of the cluster, all units within a cluster receive that treatment level, and estimands generally represent contrasts of potential outcomes at the level of the individual. We will address causal inference for cluster randomized trials. We will discuss that cluster-level randomization does not necessarily imply individual-level randomization, and formalize an assumption under which it does. In pragmatic CRTs, individuals are recruited in the study after the treatment is assigned at the cluster level, and individual recruitment can differ between treated and control clusters. What's more, data are only available among the subset of individuals that recruited. In the presence of post-randomization selection for cluster randomized trials, we will formalize causal estimands among those that recruited and in the overall population. We will link the post-randomization recruitment to covariate imbalance, and introduce the assumption of non-differential recruitment based on which we can draw causal inferences on the recruited population. Under the stronger assumption of ignorable missingness, we will show that the causal effect among the recruited control population corresponds to the causal effect among a specific tilted version of the overall population. Lastly, I will discuss sensitivity analysis for inferences on the recruited population, and the always-recruited overall population.

EO184 Room Virtual R20 RECENT DEVELOPMENTS FOR MODAL REGRESSION Chair: Lin Cong

E0331: **Spatial modal regression**
Presenter: **Tao Wang**, University of California, Riverside, United States

Modal regression with spatial data $\{(Y_i, X_i); i \in Z^N\}$ observed over a rectangular domain is proposed to be estimated by assuming that the conditional mode of the response variable Y_i given covariates X_i follows a nonparametric regression structure, defined as $m : X \mapsto m(X) = Mode(Y_i \mid X_i)$. We study the newly developed spatial modal regression by utilizing the local linear approximation augmented with shrinking bandwidths. The asymptotic normal distributions of the proposed spatial modal estimators are established, and the explicit formulas for their asymptotic biases and variances are derived under mild regularity assumptions. We also show that the targeted spatial modal regression could be used as an alternative to a nonparametric spatial mean robust regression when the data are symmetrically distributed. The asymptotic distributions for such a spatial modal-based robust estimator are derived with the appropriate choices of bandwidths. We, in the end, generalize the propounded spatial modal regression model to an additive sum of the form in order to avoid the issue of the curse of dimensionality and develop a kernel-based backfitting algorithm for estimating, where we show that the proposed spatial modal estimator of each additive component is asymptotically normal and converges at the univariate nonparametric modal optimal rate.

E0333: **A statistical learning approach to modal regression**
Presenter: **Yunlong Feng**, The State University of New York at Albany, United States

Modal regression regresses towards the conditional mode, which is another characterization of the conditional distribution in regression problems. We will consider the nonparametric modal regression problem from a statistical learning viewpoint. Technical tools that we developed to evaluate the learning performance of modal regression estimators, as well as the evaluation results, will be reported.

E0413: **Bayesian beta regression for bounded responses with unknown supports**
Presenter: **Xianzheng Huang**, University of South Carolina, United States
Co-authors: Haiming Zhou

A new Bayesian regression framework is presented for the analysis of continuous response data with support restricted to an unknown finite interval. A four-parameter beta distribution is assumed for the response conditioning on covariates, with the mean or mode depending linearly on covariates through a known link function. An informative g-prior is proposed to incorporate the prior distribution for the marginal mean or mode of the response. Byproducts of the Markov chain Monte Carlo sampling for implementing the proposed method lead to model criteria useful for model selection. Goodness-of-t of the model is assessed using Cox-Snell residual plots.

E0642: **Bootstrap inference for quantile-based modal regression**
Presenter: **Kengo Kato**, Cornell University, United States
Co-authors: David Ruppert, Tao Zhang

Uniform inference methods are developed for the conditional mode based on quantile regression. Specifically, we propose to estimate the con-

ditional mode by minimizing the derivative of the estimated conditional quantile function defined by smoothing the linear quantile regression estimator, and develop two bootstrap methods, a novel pivotal bootstrap and the nonparametric bootstrap, for our conditional mode estimator. Building on high dimensional Gaussian approximation techniques, we establish the validity of simultaneous confidence rectangles constructed from the two bootstrap methods for the conditional mode. We also extend the preceding analysis to the case where the dimension of the covariate vector is increasing with the sample size.

E0898: **Modal regression based on random forest**
Presenter: **Lin Cong**, University of California, Riverside, United States
Co-authors: Weixin Yao

Modal regression can complement the mean and quantile regressions, and provide a better central tendency measure and prediction performance when the data is skewed or heavy-tailed. Existing nonparametric modal regression has been studied from a kernel density estimation perspective. We propose to utilize a Random Forest-based quantile regression to estimate the global mode of the conditional distribution by minimizing the difference in the quotient of the conditional quantile estimators. The asymptotic property of the model estimator has been studied with its corresponding convergence rate. The validity of the proposed algorithm has been demonstrated through a set of synthetic data analyses and the performance of the algorithm on benchmark data is also compared with some other modal regression models.

EO212 Room Virtual R21 RECENT DEVELOPMENT IN HIGH-DIMENSIONAL NETWORKS Chair: Kuang-Yao Lee

E0992: **Functional differential graph estimation**
Presenter: **Y Samuel Wang**, Cornell University, United States
Co-authors: Boxin Zhao, Mladen Kolar

The problem of estimating the difference between two functional undirected graphical models with shared structures is considered. In many applications, data are naturally regarded as a vector of random functions rather than a vector of scalars. For example, electroencephalography (EEG) data are more appropriately treated as functions of time. In these problems, not only can the number of functions measured per sample be large, but each function is itself an infinite-dimensional object, making estimation of model parameters challenging. This is further complicated by the fact that the curves are usually only observed at discrete time points. We first define a functional differential graph that captures differences between two functional graphical models and formally characterize when the functional differential graph is well defined. We then propose a method, FuDGE, that directly estimates the functional differential graph without first estimating each individual graph. This is particularly beneficial in settings where the individual graphs are dense, but the differential graph is sparse. We show that FuDGE consistently estimates the functional differential graph even in a high-dimensional setting for both discretely observed and fully observed function paths.

E1098: **Functional directed acyclic graphs**
Presenter: **Bing Li**, The Pennsylvania State University, United States

A new method is introduced to estimate directed acyclic graphs from multivariate functional data, based on the notion of faithfulness that relates a directed acyclic graph with a set of conditional independence relations among the random functions. To characterize and evaluate these relations, we propose two linear operators, the conditional covariance operator and the partial correlation operator. Based on these operators, we adapt and extend the PC-algorithm to estimate the functional directed graph, so that the computation time depends on the sparsity rather than the full size of the graph. We study the asymptotic properties of the two operators, derive their uniform convergence rates, and establish the uniform consistency of the estimated graph, all of which are obtained while allowing the graph size to diverge to infinity with the sample size. We demonstrate the efficacy of our method through both simulations and an application to a time-course proteomic dataset.

E1063: **Nonparametric functional graphical models**
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 copula Gaussian distributions, which also imply linear relations among the random variables or functions on different nodes. We relax those assumptions by building our graphical model based on a new statistical object—the functional additive regression operator. By carrying out regression and neighborhood selection at the operator level, the method can capture nonlinear relations without requiring any distributional assumptions. Moreover, the method is built up using the only one-dimensional kernel, thus avoiding the curse of dimensionality from which a fully nonparametric approach often suffers, and enables us to work with large-scale networks. We derive error bounds 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. We demonstrate the efficacy of our method by both simulations and analysis of an electroencephalography dataset.

E1002: **Fast variational inference for joint mixed sparse graphical models**
Presenter: **Yuping Zhang**, University of Connecticut, United States
Co-authors: Qingyang Liu

A statistical learning framework is presented for multiple mixed graphical models via a penalized approximate likelihood estimation. We describe a fast algorithm for variational maximum likelihood inference, which takes advantage of the log-determinant relaxation. We identify a necessary and sufficient condition to discover the connected components in the solution. We then employ a divide-and-conquer approach to the joint structural inference problem for multiple related large sparse networks.

E1410: **Conditional independence testing for categorical data**
Presenter: **Harvey Klyne**, University of Cambridge, United Kingdom
Co-authors: Rajen D Shah

The focus is on the problem of testing whether X and Y, at least one of which is categorical, are conditionally independent, given a random vector Z. It is known that when components of Z are continuous, no uniformly valid conditional independence test can hold power against any alternative, and so non-trivial tests can only hope to maintain the level over subsets of the null. We propose a test statistic based on the residuals from regressing a one-hot coding of the categorical data onto Z that provides type I error control whenever the prediction errors decay sufficiently fast. The required rates are slow enough to accommodate settings where Z is high-dimensional, and when the regression functions are nonparametric, these scenarios are not being covered by the standard log-linear analysis, for example. For cases where the number of levels of the categorical data is moderate to large, we propose an algorithm to optimally aggregate levels when performing the test that also maintains type I error control under similar conditions.

EO332 Room Virtual R22 ROBUSTNESS AND DATA ANALYSIS Chair: Graciela Boente

E0265: **Robust optimal estimation of location from discretely sampled functional data**
Presenter: **Ioannis Kalogridis**, KU Leuven, Belgium
Co-authors: Stefan Van Aelst

Estimating location is a central problem in functional data analysis. Yet, most current estimation procedures either unrealistically assume completely observed trajectories or lack robustness with respect to the many kinds of anomalies one can encounter in the functional setting. To remedy these

deficiencies, we introduce a class of optimal robust location estimators based on discretely sampled functional data. The proposed method is based on M-type smoothing spline estimation with repeated measurements and is suitable for both commonly and independently observed trajectories that are subject to measurement error. We show that under suitable assumptions, the proposed family of estimators is minimax rate optimal both for commonly and independently observed trajectories. We illustrate its highly competitive performance and practical usefulness in a Monte-Carlo study and a real-data example involving recent Covid-19 data.

E0841: **Real-time discriminant analysis in the presence of label and measurement noise**
Presenter: **Mia Hubert**, KU Leuven, Belgium
Co-authors: Iwein Vranckx, Jakob Raymaekers, Bart de Ketelaere, Peter Rousseeuw

Quadratic discriminant analysis (QDA) is a widely used classification technique. Based on a training dataset, each class in the data is characterized by an estimate of its center and shape, which can then be used to assign unseen observations to one of the classes. The traditional QDA rule relies on the empirical mean and covariance matrix. Unfortunately, these estimators are sensitive to label and measurement noise which often impairs the models predictive ability. Robust estimators of location and scatter are resistant to this type of contamination. However, they have a prohibitive computational cost for large scale industrial experiments. We present a novel QDA method based on a recent real-time robust algorithm. We additionally integrate an anomaly detection step to classify the most atypical observations into a separate class of outliers. Finally, we introduce the class map. Its goal is to visualize aspects of the classification results to obtain insight into the data.

E1147: **Robust boosting for regression problems**
Presenter: **Matias Salibian-Barrera**, The University of British Columbia, Canada
Co-authors: Xiaomeng Ju

Gradient boosting algorithms construct a regression predictor using a linear combination of base learners. Boosting also offers a family of non-parametric regression estimators that are scalable to applications with many explanatory variables. The robust boosting algorithm is based on a two-stage approach, similar to what is done for robust linear regression: it first minimizes a robust residual scale estimator, and then improves it by optimizing a bounded loss function. Unlike previous robust boosting proposals, our approach does not require computing an ad-hoc residual scale estimator in each boosting iteration. A robust variable importance measure can also be calculated via a permutation procedure. Through simulation studies and several data analyses show that, when no atypical observations are present, the robust boosting approach works as well as the standard gradient boosting with a squared loss. Furthermore, when the data contain outliers, the robust boosting estimator outperforms the alternatives in terms of prediction error and variable selection accuracy.

E0268: **Robust testing to compare regression curves**
Presenter: **Juan-Carlos Pardo-Fernandez**, Universidade de Vigo, Spain
Co-authors: Graciela Boente

The problem of testing for the equality of regression curves against general alternatives in a fully nonparametric setting is considered. A test statistic based on an L_2-distance between empirical characteristic functions of residuals is considered. To protect against atypical observations, the residuals are obtained by using robust estimates of the regression functions. The asymptotic distribution of the test statistic is analysed, and a small Monte Carlo study is performed to investigate the finite sample behaviour of the proposed test.

E0336: **Partially linear single-index models: A robust approach**
Presenter: **Ana Maria Bianco**, Universidad de Buenos Aires, Argentina
Co-authors: Maria Florencia Statti

When using fully nonparametric models, practitioners often face the curse of dimensionality. In this context, dimension reduction becomes a relevant issue. Partially linear single-index models are a good strategy to reduce dimension and capture nonlinear trends simultaneously. These models are a reasonable trade-off between the fully parametric and fully non-parametric approaches. We propose a robust two-stage estimation procedure of the parametric and nonparametric components of the model when the scale parameter is unknown. We study the consistency of the estimators and derive the asymptotic distribution of the linear and single index parameters. A simulation study is performed, and an application to a real dataset is illustrated. We also explore the finite sample properties of a Wald-type test to check hypotheses that involved the linear parameter.

EO414 Room Virtual R23 MODERN APPROACHES TO BIOMEDICAL DATA ANALYSIS Chair: Sunyoung Shin

E0477: **Model-assisted uniformly honest inference for optimal treatment regimes in high dimension**
Presenter: **Yunan Wu**, The University of Texas at Dallas, United States
Co-authors: Lan Wang

New tools are developed to quantify uncertainty in optimal decision-making and to gain insight into which variables one should collect information about, given the potential cost of measuring a large number of variables. We investigate simultaneous inference to determine if a group of variables is relevant for estimating an optimal decision rule in a high-dimensional semiparametric framework. The unknown link function permits flexible modeling of the interactions between the treatment and the covariates, but leads to nonconvex estimation in high dimension and imposes significant challenges for inference. We first establish that a local restricted strong convexity condition holds with high probability and that any feasible local sparse solution of the estimation problem can achieve the near-oracle estimation error bound. We further rigorously verify that a wild bootstrap procedure based on a debiased version of the local solution can provide asymptotically honest uniform inference for the effect of a group of variables on optimal decision making. The advantage of honest inference is that it does not require the initial estimator to achieve perfect model selection and does not require the zero and nonzero effects to be well-separated. We also propose an efficient algorithm for estimation. Our simulations suggest satisfactory performance. An example from a diabetes study illustrates the real application.

E0561: **Semiparametric Gumbel regression model for analyzing longitudinal data with non-normal tails**
Presenter: **Noorie Hyun**, Medical College of Wisconsin, United States
Co-authors: David Couper, Donglin Zeng

Abnormal longitudinal values in biomarkers can be a sign of abnormal status or disease. Identifying new biomarkers for early and efficient disease detection is crucial for disease prevention. Compared to the majority of the healthy general population, abnormal values are located within the tails of the biomarker distribution. Thus, parametric regression models that accommodate abnormal values in biomarkers can be better for detecting the association between biomarkers and disease. We propose semiparametric Gumbel regression models for (1) longitudinal continuous biomarker outcomes, (2) flexibly modeling the time-effect on the outcome and (3) accounting for the measurement error in biomarker measurements. We adopted the EM algorithm in combination with a two-dimensional grid search to estimate regression parameters and a function of time-effect. We proposed an efficient asymptotic variance estimator for regression parameter estimates. The proposed estimator is asymptotically unbiased in both theory and simulation studies. We applied the proposed model and two other models to investigate associations between fasting blood glucose biomarkers and potential risk factors from a diabetes ancillary study to the Atherosclerosis Risk in Communities (ARIC) study. The real data application was illustrated by fitting the proposed regression model and by graphically evaluating the goodness-of-fit value.

E0671: **Transfer learning for cognitive reserve quantification**
Presenter: **Seonjoo Lee**, Columbia University/New York State Psychiatric Institute, United States
Co-authors: Xi Zhu, Yi Liu, Christian Habeck, Yaakov Stern

Cognitive reserve has been introduced to explain individual differences in susceptibility to cognitive or functional impairment in the presence of age or pathology. We developed a deep learning model to quantify the CR as residual variance in memory performance using the structural MRI data from a lifespan healthy cohort. The generalizability of the sMRI-based deep learning model was tested in two independent healthy and Alzheimer cohorts using a transfer learning framework. Structural MRIs were collected from three cohorts: 495 healthy adults from RANN, 620 healthy participants (age 36-100) from lifespan Human Connectome Project Aging (HCPA), and 941 subjects from Alzheimer's Disease Neuroimaging Initiative (ADNI). Cognitive reserve was quantified by residuals which subtract the predicted memory from the true memory. Cascade neural network (CNN) models were used to train the RANN dataset for memory prediction. The CNN model trained on the RANN dataset exhibited a strong linear correlation between true and predicted memory based on the chosen T1 cortical thickness and volume predictors. In addition, the model generated from healthy lifespan data (RANN) was able to generalize to independent healthy lifespan data (HCPA) and older demented participants (ADNI) across different scanner types. The estimated CR was correlated with CR proxies such as education and IQ across all three datasets.

E0688: **On estimation and selection for semiparametric models in meta-analysis**
Presenter: **Sunyoung Shin**, University of Texas at Dallas, United States

Combining large-scale datasets of multiple studies is a valuable approach to fully utilizing the collected data. However, such studies often have privacy policies or data transfer issues that prevent individual-level data sharing. A meta-analysis combines large-scale datasets using compressed information in summary statistics without requiring individual-level data. We develop a general likelihood theory on meta-analysis with semiparametric models. The theoretical framework embraces meta-analysis of studies with different observation schemes that generate various data types. We propose a method of meta-estimation and selection based on summary statistics. The resulting estimator has desirable asymptotic properties under mild assumptions. The superior performance and practical utility of the proposed method are demonstrated through numerical studies.

E1087: **Bayesian analysis of longitudinal dyadic/multiple outcome data with informative missing data**
Presenter: **Jaeil Ahn**, Georgetown University, United States

Analysis of longitudinal dyadic/multiple outcomes with missing data is challenging due to the complicated correlations within and between dyads/multiple outcomes, as well as non-ignorable missing data. We will introduce a Bayesian mixed-effects hybrid model to analyze longitudinal dyadic data with non-ignorable dropouts/intermittent missingness. To address this, we factorize the joint distribution of the measurement, random effects, and dropout processes into three components. The proposed model accounts for the dyadic interplay using the concept of actor and partner effects as well as dyad-specific random effects. We evaluate the performance of the proposed methods using a simulation study, and apply our method to longitudinal dyadic datasets that arose from a prostate cancer trial. We will introduce a Bayesian mixed-effects selection model to analyze the multivariate quality of life data with non-ignorable missing data. Compared to the first model, we first describe the overall/local effects of predictors on outcomes simultaneously and then incorporate a variable selection feature in the missing data mechanism to evaluate the impact of potentially moderate to high dimensional outcomes on missing data mechanisms.

EO230 Room Virtual R24 REINFORCEMENT LEARNING WITH APPLICATIONS TO PRECISION MEDICINE Chair: Hengrui Cai

E0222: **Doubly robust interval estimation for optimal policy evaluation in online learning**
Presenter: **Hengrui Cai**, North Carolina State University, United States
Co-authors: Ye Shen, Rui Song

Evaluating the performance of an ongoing policy plays a vital role in many areas such as medicine and economics, to provide crucial instruction on the early-stop of the online experiment and timely feedback from the environment. Policy evaluation in online learning thus attracts increasing attention by inferring the mean outcome of the optimal policy (i.e., the value) in real-time. Yet, such a problem is particularly challenging due to the dependent data generated in the online environment, the unknown optimal policy, and the complex exploration and exploitation trade-off in the adaptive experiment. We aim to overcome these difficulties in policy evaluation for online learning. We explicitly derive the probability of exploration that quantifies the probability of exploring the non-optimal actions under commonly used bandit algorithms. We use this probability to conduct valid inference on the online conditional mean estimator under each action and develop the doubly robust interval estimation (DREAM) method to infer the value under the estimated optimal policy in online learning. The proposed value estimator provides double protection on the consistency and is asymptotically normal with a Wald-type confidence interval provided. Extensive simulations and real data applications are conducted to demonstrate the empirical validity of the proposed DREAM method.

E0244: **Learning individualized treatment rules for a target population**
Presenter: **Guanhua Chen**, University of Wisconsin-Madison, United States

Learning individualized treatment rules (ITRs) is an important topic in precision medicine. Current literature mainly focuses on deriving ITRs from a single source population. We consider the observational data setting when the source population differs from a target population of interest. We assume subject covariates are available from both populations, but treatment and outcome data are only available from the source population. Although adjusting for differences between source and target populations can potentially lead to an improved ITR for the target population, it can substantially increase the variability in ITR estimation. To address this dilemma, we develop a weighting framework that aims to tailor an ITR for a given target population and protect against high variability due to superfluous covariate shift adjustments. Our method seeks covariate balance over a nonparametric function class characterized by a reproducing kernel Hilbert space and can improve many ITR learning methods that rely on weights. We show that the proposed method encompasses importance weights and the so-called overlap weights as two extreme cases, allowing for a better bias-variance trade-off in between. Numerical examples demonstrate that the use of our weighting method can greatly improve ITR estimation for the target population compared with other weighting methods.

E0368: **Post-selection inference for individualized treatment rules**
Presenter: **Ashkan Ertefaie**, University of Rochester, United States
Co-authors: Robert Strawderman, Jeremiah Jones

Constructing an optimal treatment regime become complex when there is a large number of prognostic factors, such as patients genetic information, demographic characteristics, medical history over time. Existing methods only focus on selecting the important variables for the decision-making process and fall short in providing inference for the selected model. We fill this gap by leveraging the conditional selective inference methodology. We show that the proposed method is asymptotically valid given certain rate assumptions in semiparametric regression.

E0557: **Statistical efficient batch policy learning in average reward markov decision processes**
Presenter: **Zhengling Qi**, The George Washington University, United States

The batch (off-line) reinforcement learning problem in infinite horizon Markov Decision Processes is discussed. Motivated by mobile health studies, we focus on learning a policy that maximizes the long-term average reward. Given limited pre-collected data, we propose a doubly robust estimator for the average reward and show that it achieves statistical efficiency bound. We then develop an optimization algorithm to compute the optimal policy in a parametrized stochastic policy class. The performance of the estimated policy is measured by the difference between the optimal average reward in the policy class and the average reward of the estimated policy. Under some technical conditions, we establish a strong finite-sample regret guarantee in terms of total decision points, demonstrating that our proposed method can efficiently break the curse of horizon. Finally, the performance of the proposed method is illustrated by simulation studies.

E1387: **Efficient learning and evaluation of individualized treatment rules under data fusion**
Presenter: **Alex Luedtke**, University of Washington, United States
The aim is to fuse data from multiple sources together to learn and make inferences about generic smooth summaries of an individualized treatment rule, such as its mean outcome or the proportion of people that it recommends treating. Previous works have studied the estimation of a variety of parameters in similar data fusion settings, including in the estimation of the average treatment effect, optimal treatment rule, or average reward, with the majority of them merging one historical dataset with covariates, actions, and rewards and one dataset of the same covariates. We consider the general case where multiple datasets align with different parts of the distribution of the target population, for example, the conditional distribution of the reward given actions and covariates. We then examine potential gains in efficiency that can arise from fusing these datasets together in a single analysis, which we characterize by a reduction in the semiparametric efficiency bound. In numerical experiments, we show marked improvements in efficiency from using our proposed estimators compared to their natural alternatives.

EO537 Room Virtual R25 CAUSAL INFERENCE CHALLENGES IN HEALTH POLICY DECISION MAKING Chair: Nandita Mitra

E0178: **Within- versus between-market comparison units for diff-in-diff**
Presenter: **Laura Hatfield**, Harvard Medical School, United States
Difference-in-differences (diff-in-diff) is a popular method for causal inference in observational settings. It requires outcomes of units exposed to the intervention (treated) and units not exposed to the intervention (control), both before and after the intervention. The key causal assumption is that pre- to post-intervention changes in the outcome of the treated and control groups would have been the same in the absence of treatment. In some settings, we may be able to use comparison units from within the same market/region as the treated units. This could strengthen the plausibility of the causal assumption (because control units are subject to the same market forces), but may raise the concern that treated units are on different trajectories or that too many units were treated to leave a good pool of potential controls. We address the question: under what conditions are out-of-market controls preferable to within-market controls? We use simulations and real data analysis to show the combinations of within- and between-market variability and systematic selection forces that lead us to prefer one or the other. We give special attention to the impact of COVID-19 on policy evaluations and the challenges of the time- and space-varying impacts of the pandemic.

E0908: **Extending synthetic control methods to evaluate the effectiveness of global maternal health programs**
Presenter: **Isabel Fulcher**, Harvard Medical School, United States
Community health worker (CHW) programs are commonly used to improve access to prenatal, postnatal, and neonatal care in low- and middle-income countries by connecting families to facility-based care. In CHW-led programs operating at scale, rich individual-level data is often collected among persons receiving the intervention as a product of program implementation; however, data is often not collected at baseline or among a comparison group post-intervention. Recent advances in causal inference provide methods for emulating a trial when randomization is not possible or, in the case of synthetic control methods, when there is not an obvious comparison group. We present an extension of synthetic control methods to account for scenarios when data is not collected among the intervention group at baseline. We apply these methods to a large maternal health program in Zanzibar, Tanzania, utilizing data from the program and publicly available Demographic and Health Surveys data.

E1082: **Identifying optimally cost-effective regimes with a Q-learning approach**
Presenter: **Nicholas Illenberger**, University of Pennsylvania, United States
Health policy decisions regarding patient treatment strategies require consideration of both treatment effectiveness and cost. Optimizing treatment rules with respect to effectiveness may result in prohibitively expensive strategies; on the other hand, optimizing with respect to costs may result in poor patient outcomes. We propose a two-step approach for identifying an optimally cost-effective and interpretable dynamic treatment regime. First, we develop a combined Q-learning and policy-search approach to estimate an optimal list-based regime under a constraint on expected treatment costs. Second, we propose an iterative procedure to select an optimally cost-effective regime from a set of candidate regimes corresponding to different cost constraints. Our approach can estimate optimal regimes in the presence of commonly encountered challenges including time-varying confounding and correlated outcomes. Through simulation studies, we illustrate the validity of estimated optimal treatment regimes and examine operating characteristics under flexible modeling approaches. Using data from an observational cancer database, we apply our methodology to evaluate optimally cost-effective treatment strategies for assigning adjuvant radiation and chemotherapy to endometrial cancer patients.

E1138: **(Counterfactually) fair and accurate risk assessment for healthcare decision making**
Presenter: **Alan Mishler**, J.P. Morgan Chase, United States
Algorithmic tools are increasingly used in healthcare settings to identify high-risk patients and inform treatment decisions, but these tools may perform differently across different racial groups (for example), leading to concerns about fairness. Although many methods exist for developing fair predictors, most such methods are concerned with observable outcomes, such as actual patient survival, which depends both on a patient's health and on the treatment they receive. By contrast, the accuracy and fairness properties of risk assessment tools are often most sensibly understood in terms of counterfactual outcomes: how a patient would fare if given, or not given, a particular treatment, irrespective of the treatment they actually receive. We (1) illustrate how a reliance on observable outcomes in risk assessment tools can worsen these outcomes, leading for example to higher patient mortality; and (2) describe a set of methods for building predictors that are both fair and accurate with respect to relevant counterfactual outcomes. These methods accommodate a wide range of fairness criteria, and they facilitate computationally efficient exploration of fairness-accuracy and fairness-fairness tradeoffs. In some cases, multiple unfairness measures can be simultaneously minimized with little cost in accuracy relative to a benchmark model.

E1723: **Discussant**
Presenter: **Alan Hubbard**, University of California, Berkeley, United States
The purpose is to discuss what the talks imply about causal inference used in health policy decisions.

EO683 Room Virtual R26 DATA INTEGRATION METHODS AND APPLICATIONS Chair: Hai Shu

E0180: **Simultaneous clustering and estimation of networks in multiple graphical models**
Presenter: **Gen Li**, University of Michigan Ann Arbor, United States
The standard Gaussian graphical models have been widely used to investigate the dependency structure among variables in a single population. As it is increasingly common to collect data from multiple heterogeneous populations, multi-layered networks have become prevalent. We consider the simultaneous clustering and estimation of multiple graphical models. We build upon the Gaussian graphical models and utilize a sparse tensor decomposition approach to simultaneously cluster populations and estimate the underlying network structures among variables in each population. A penalized likelihood method is used to devise an alternating direction method of multipliers algorithm to estimate model parameters. We demonstrate the efficacy of the proposed method with comprehensive simulation studies. The application to the GTEx multi-tissue gene expression data provides important insights into tissue clustering and gene co-expression patterns in different tissues.

E0374: **Data integration to improve prediction of human complex traits and diseases**
Presenter: **Bingxin Zhao**, Purdue University, United States
Co-authors: Hongtu Zhu
One ultimate goal in biomedical studies is to develop prediction models for complex traits and diseases. We use large-scale datasets to showcase

some recent real data applications to predict complex traits and diseases (such as fluid intelligence and heart diseases). We integrate multiple data resources, including the common genetic variants from high-dimensional genotyping data, gene expression data, exome data, biomarkers, and multi-modality imaging traits. We illustrate the clinical achievements of current data integration methods and also highlight the existing challenges and opportunities.

E0547: **Mixture of shape-on-scalar regression models:going beyond prealigned non-Euclidean responses**
Presenter: **Chao Huang**, Florida State University, United States

Due to the wide applications of shape data analysis in medical imaging, computer vision, and many other fields, it is of great interest to cluster objects and recovers the underlying sub-group structure according to their shapes and covariates in Euclidean space (e.g., age and diagnostic status). However, this clustering task faces four challenges including (i) non-Euclidean space, (ii) misalignment of shapes due to pre-processing steps and imaging heterogeneity, (iii) complex spatial correlation structure, and (iv) geodesic variation associated with some covariates. In order to address these challenges, we propose a mixture of geodesic factor regression models (M-GeoFARM). In each cluster, a geodesic regression structure including covariates of interest and alignment step is established along with the RiemannianGaussian distribution in the pre-shape space, and a latent factor model is built in the tangent space. In addition, a Monte Carlo EM algorithm is provided for the parameter estimation procedure. Finally, both simulation studies and real data analysis are conducted to compare the clustering performance of M-GeoFARM with other existing methods.

E0563: **Improving personalized causal inference with information borrowed from heterogeneous data sources**
Presenter: **Xiaoqing Tan**, University of Pittsburgh, United States
Co-authors: Lu Tang

Individualized causal inference, ranging from personalized medicine to customized marketing advertisement, has remained a hot topic. However, due to the limited sample size in a single study, estimating treatment effects or optimal treatment rules is often challenging. We propose a tree-based model averaging framework to improve the estimation efficiency of conditional average treatment effects (CATE) and optimal decision rules concerning the population of a targeted research site by leveraging models derived from potentially heterogeneous populations of other sites, but without them sharing individual-level data. To our best knowledge, there is no established model averaging approach for distributed data with a focus on improving the estimation of treatment effects. Drawing on a multi-hospital electronic health records network, we develop an efficient and interpretable tree-based ensemble of personalized treatment effect estimators to join results across hospital sites, while actively modeling for the heterogeneity in data sources through site partitioning. The efficiency of this approach is demonstrated by a study of causal effects of oxygen saturation on hospital mortality and backed up by comprehensive numerical results.

E0349: **Common and distinctive pattern analysis between high-dimensional data sets**
Presenter: **Hai Shu**, New York University, United States

A representative model in integrative analysis of two high-dimensional correlated datasets is to decompose each data matrix into a low-rank common matrix generated by latent factors shared across datasets, a low-rank distinctive matrix corresponding to each dataset, and an additive noise matrix. Existing decomposition methods claim that their common matrices capture the common pattern of the two datasets. However, their so-called common pattern only denotes the common latent factors but ignores the common pattern between the two coefficient matrices of these common latent factors. We propose a new unsupervised learning method, called the common and distinctive pattern analysis (CDPA), which appropriately defines the two types of data patterns by further incorporating the common and distinctive patterns of the coefficient matrices. A consistent estimation approach is developed for high-dimensional settings, and shows reasonably good finite-sample performance in simulations. Our simulation studies and real data analysis corroborate that the proposed CDPA can provide better characterization of common and distinctive patterns and thereby benefit data mining.

EO772 Room Virtual R27 RECENT ADVANCES IN DATA PRIVACY Chair: Matthew Reimherr

E1461: **Recent advances in private synthetic data generation**
Presenter: **Steven Wu**, Carnegie Mellon University, United States

The focus is on differentially private synthetic data—a privatized version of the dataset that consists of fake data records and that approximates the real dataset on important statistical properties of interest. We will present our recent results on private synthetic data that leverage practical optimization heuristics to circumvent the computational bottleneck in existing work. The techniques are motivated by a modular, game-theoretic framework, which can flexibly work with methods such as integer program solvers and deep generative models.

E0554: **Mean estimation with user-level privacy under data heterogeneity**
Presenter: **Rachel Cummings**, Columbia University, United States

A key challenge for data analysis in the federated setting is that user data is heterogeneous, i.e., it cannot be assumed to be sampled from the same distribution. Further, in practice, different users may possess a vastly different number of samples. We propose a simple model of heterogeneous user data that differs in both distribution and quantity of data, and we provide a method for estimating the population level meanwhile preserving user-level differential privacy. We demonstrate the asymptotic optimality of the estimator within a natural class of private estimators and also prove general lower bounds on the error achievable in our problem. In particular, while the optimal non-private estimator can be shown to be linear, we show that privacy constrains us to use a non-linear estimator.

E0593: **High-dimensional differentially-private em algorithm: Methods and near-optimal statistical guarantees**
Presenter: **Linjun Zhang**, Rutgers University, United States

A general framework is developed to design differentially private expectation-maximization (EM) algorithms in high-dimensional latent variable models, based on the noisy iterative hard-thresholding. We derive the statistical guarantees of the proposed framework and apply it to three specific models: Gaussian mixture, mixture of regression, and regression with missing covariates. In each model, we establish the near-optimal rate of convergence with differential privacy constraints, and show the proposed algorithm is minimax rate optimal up to logarithm factors. The technical tools developed for the high-dimensional setting are then extended to the classic low-dimensional latent variable models, and we propose a near rate-optimal EM algorithm with differential privacy guarantees in this setting. Simulation studies and real data analysis are conducted to support our results.

E0743: **Differential privacy over Riemannian manifolds**
Presenter: **Carlos Soto**, The Pennsylvania State University, United States
Co-authors: Matthew Reimherr, Karthik Bharath

The problem of releasing a differentially private statistical summary that resides on a Riemannian manifold is considered. It presents an extension of the Laplace, or K-norm, a mechanism that utilizes intrinsic distances and volumes while specifically considering the case where the summary is the Fréchet mean. The mechanism is shown to be rate optimal and depends only on the dimension of the manifold, not on the dimension of an ambient space, while also showing that ignoring the manifold structure can decrease the utility of the privatized summary. The proposed framework is illustrated in two examples of particular interest in statistics: the space of symmetric positive definite matrices, which is used for covariance matrices, and the sphere, which can be used as a space for modeling discrete distributions.

E1048: **Canonical noise distributions and private hypothesis tests**
Presenter: **Jordan Awan**, Purdue University, United States
Co-authors: Salil Vadhan

f-DP has recently been proposed as a generalization of classical definitions of differential privacy allowing a lossless analysis of composition, post-processing, and privacy amplification via subsampling. In the setting of f-DP, we propose the concept *canonical noise distribution* (CND) which captures whether an additive privacy mechanism is appropriately tailored for a given f, and give a construction that produces a CND given an arbitrary tradeoff function f. We show that private hypothesis tests are intimately related to CNDs, allowing for the release of private p-values at no additional privacy cost as well as the construction of uniformly most powerful (UMP) tests for binary data. We apply our techniques to the problem of difference of proportions testing, and construct a UMP unbiased "semi-private" test that upper bounds the performance of any DP test. Using this as a benchmark we propose a private test, based on the inversion of characteristic functions, which allows for optimal inference for the two population parameters and is nearly as powerful as the semi-private UMPU. When specialized to the case of $(\varepsilon,0)$-DP, we show empirically that our proposed test is more powerful than any $(\varepsilon/\sqrt{2})$-DP test and has more accurate type I errors than the classic normal approximation test.

EO822 Room Virtual R28 SMALL AREA ESTIMATION AND PUBLIC STATISTICS Chair: Maria Guadarram Sanz

E0509: **Small area estimation in the presence of data masking**
Presenter: **Nikos Tzavidis**, University of Southampton, United Kingdom

The production of official statistics is explored, in particular, estimation for small geographic areas in the presence of data masking. The main focus will be on estimation when the response variable is grouped due to concerns about data confidentiality or survey response burden. Reporting data in groups (bands) is a mechanism that is employed, for example, in surveys collecting information on income. Methodology that enables fitting a random-effects model when the dependent variable is grouped will be outlined. Model parameters are then used for small area prediction of finite population parameters. Model fitting is based on the use of a stochastic EM (SEM) algorithm. Since the SEM algorithm relies on Gaussian assumptions, adaptive transformations are developed for handling departures from normality. The estimation of the mean squared error of the small area parameters is facilitated by a parametric bootstrap that captures the additional uncertainty due to the censoring mechanism and the use of transformations. The presentation will also briefly discuss the production of official statistics (a) when geographical coordinates are aggregated and (b) when data is geographically displaced.

E0933: **Single source small area estimation**
Presenter: **Jan Pablo Burgard**, Trier University, Germany
Co-authors: Domingo Morales, Joscha Krause

Regional indicators are important for business decisions and policymaking. Most regional information is not collected in registers, but gathered in surveys. Due to disclosure reasons and small sample sizes, only aggregated, imprecise totals or mean estimates are provided. Small area models, that aim to improve the precision of the regional estimates, must explicitly account for data uncertainty to allow for reliable results. This can be achieved via measurement error models that introduce distribution assumptions on the noisy data. However, these methods usually require target and explanatory variable errors to be independent. This does not hold when data for both have been estimated from the same survey, which is sometimes the case in official statistics or special purpose surveys. If not accounted for, prevalence estimates can be severely biased. We propose a new measurement error model for regional prevalence estimation that is suitable for settings where target and explanatory variable errors are dependent. We derive the empirical best predictors and demonstrate mean-squared error estimation. A maximum likelihood approach for model parameter estimation is presented. Simulation experiments are conducted to prove the effectiveness of the method. An application to regional hypertension prevalence estimation in Germany is provided.

E0622: **Inference for big data assisted by small area methods**
Presenter: **Gaia Bertarelli**, Sant'Anna school of Advanced Studies, Italy
Co-authors: Francesco Schirripa Spagnolo, Stefano Marchetti, Nicola Salvati, Monica Pratesi

Nowadays, the availability of a huge amount of data produced by a wide range of new technologies, so-called big data, is increasing. Their availability to unprecedented spatial detail represents an opportunity in the context of Small Area Estimation (SAE) to infer some characteristics for very small domains. However, data obtainable from big data sources are often the result of a non-probability sampling process and adjusting for the selection bias is an important practical problem. We propose a novel method of reducing the selection bias associated with the big data source in SAE. The approach is based on data integration and onto the combination of a big data sample and a probability sample. We are interested in the estimation of the population mean of a target variable in each small area of interest. We assume the target variable is available from the big data sources, while auxiliary variables are also available from survey samples. Because of the selection bias, the sample mean of the target variable calculated using the big data is biased and by incorporating the auxiliary information from an external source, we can reduce the selection bias. We develop doubly robust estimators with their MSEs by using SAE models with area-specific effects. These models are implemented to obtain the area estimator from the sample data and the parameters of the propensity score for the big data sample.

E0961: **Official statistics based on the Dutch health survey during the COVID-19 pandemic**
Presenter: **Jan van den Brakel**, Statistics Netherlands, Netherlands

The Dutch Health Survey (DHS), conducted by Statistics Netherlands, is designed to produce reliable direct estimates about health-related themes at an annual frequency. Data collection is based on a combination of web interviewing (CAWI) and face-to-face interviewing (CAPI). During the COVID-19 lockdown, CAPI partially stopped, which results in a sudden change in measurement and selection effects in the survey outcomes. Furthermore, the production of annual data about the effect of CODID-19 on health-related themes with a delay of about one year compromises the relevance of this survey. The sample size of the DHS does not allow the production of figures for shorter reference periods. Both issues are solved by developing a bivariate structural time series model to estimate quarterly figures for the most important key variables. The input series are quarterly direct estimates based on the complete response of CAPI and CAWI and a series of direct estimates based on the CAWI response only. During the lockdown, the direct estimates for the complete response are missing and the time series model provides an optimal nowcast for this figure. The model is also used as a form of small area estimation that borrows sample information observed in previous reference periods. In this way, timely and relevant statistics that describe the effects of the corona crisis on the development of health, medical contacts, lifestyle and preventive behaviour in the Netherlands are published.

E1706: **SPREE estimation of the number of disabled people in terms of economic activity**
Presenter: **Marcin Szymkowiak**, Poznan University of Economics and Business and Statistical Office in Poznan, Poland

The Labour Force Survey (LFS) is the basic source of information published by Statistics Poland about the situation regarding the economic activity of the population, i.e. the fact of being employed, unemployed, or economically inactive, both for the country as a whole and at the regional level (NUTS 2 - province). It causes that estimates for lower levels of the territorial division or more detailed domains to have not been published so far. This also applies to a very important phenomenon of disability, especially in terms of the economic activity of disabled people in the labour market. So far, only basic characteristics on the population of employed, unemployed and economically inactive disabled persons based on the legal criterion from the LFS have been published. The main aim is to show chosen results of estimation of disability in Poland at a lower level of aggregation than the one used so far, that is at the level of a province with additional cross-classification including place of living using chosen small area estimation methods, i.e. the Structure Preserving Estimation (SPREE) and its generalization based on GLSM and GLSMM estimators.

By choosing and combining data from LFS 2011-2020 and Census in Poland we can obtain results with adequate precision. Territorial analysis of the scope of disability in Poland at NUTS 2 level with additional breakages will be also presented in detail.

EO844 Room Virtual R29 ADAPTIVE METHODS FOR COMPLEX HIGH DIMENSIONAL TIME SERIES ANALYSIS Chair: Scott Bruce

E0829: **Causality in multivariate time series with mixed components**
Presenter: **Wagner Barreto-Souza**, King Abdullah University of Science and Technology, Saudi Arabia
Co-authors: Hernando Ombao

The Granger causality is investigated in a general class of multivariate time series models that allow for a wide range of mixed components including: non-negative, count, bounded, binary, and real-valued time series. We explore the asymptotic properties of the proposed Granger causality test. Simulated and real data analyses are presented to illustrate the potential for the practice of the developed methodology.

E1375: **A frequency-domain multivariate linear model for analyzing multiple time series and covariates**
Presenter: **Zeda Li**, City University of New York, United States
Co-authors: Yuexiao Dong

A frequency-domain multivariate linear model is proposed to study the association between covariates and the second-order power spectra of multiple time series. A random Cramer representation, where power spectra are assumed to be random functions that are correlated with the covariates, is used as a joint model for collections of time series and covariates. Each subject-specific time series is represented by a set of cepstral coefficients, allowing the proposed model to capture frequency patterns of the time series parsimoniously. A multivariate linear model concerning the cepstral coefficients and covariates is then constructed to provide flexible yet interpretable measures of association between power spectra and covariates. The parameters of the multivariate linear model can be estimated by the envelope estimator, which provides a tool for dimension reduction and is more robust than the ordinary least squares estimator when the covariates are highly correlated. Empirical performance is evaluated in simulation studies and illustrated through a study of gait variability in young children.

E1495: **Use of wavelet based spectra for early detection of ovarian cancer**
Presenter: **Dixon Vimalajeewa**, Texas A&M University, United States
Co-authors: Scott Bruce, Brani Vidakovic

Ovarian cancer presents at a late clinical-stage so that early detection of this cancer is essential for improving the survival rate. Serum mass spectrometry data collected from ovarian cancer and non-cancer patients are commonly used early diagnosis of ovarian cancer. Previous studies have mostly considered only some specific features from the spectra in detecting the presence of cancer. However, the whole spectra are accounted for and a new modality is discussed by using wavelet analysis. Wavelet analysis is a popular signal processing tool that transforms a signal into a set of coefficients representing the signal's nature at different locations and times. Wavelet spectrum is formed by using these coefficients, and the spectral slope of the wavelet spectra is used to measure the signal's regularity. The study discusses signal classification based on variability in signals regularity and its usability in the early detection of ovarian cancer. Spectral slopes of ovarian cancer spectra are computed by using wavelet spectra formed through the standard and distance covariance-based methods. Those spectral slopes are then fed into three classification algorithms, logistic regression, SVM, and KNN. Finally, the contribution of regularity in ovarian mass spectroscopy data on detecting cancer from non-cancer data is assessed with respect to the spectral slope computing methods and the classification algorithms by using correct classification rate, sensitivity, and specificity.

E1432: **Inference for nonhomogeneous Bellman-Harris processes with applications to transportation analytics**
Presenter: **Pramita Bagchi**, George Mason University, United States
Co-authors: Anand Vidyashankar

Vehicle sharing systems such as bikeshare, scooter share, and car share are undergoing continuous improvements to increase the adoption of public transportation. The difficulties of the last-mile problem can only be solved if the system is convenient, reliable, and cost-effective. Big data methods such as learning algorithms combined with optimization techniques are increasingly used to understand the mobility patterns of customers and the demand for vehicles within a transportation system, yielding empirical solutions. While these methods facilitate some data-driven decision-making, they tend to have limited applicability due to the inherent ad hoc nature of the procedures. We develop an alternative approach, based on a non-homogeneous age-dependent branching process, that incorporates differential dynamics of the vehicle usage across time and "stations". We then cast various scientific questions as inferential questions concerning the parameters of the model. We address the resulting inferential issues using rigorous statistical and computational approaches. For this reason, we establish central limit theorems concerning functionals of the non-homogeneous age-dependent branching processes and use them to develop algorithms for real-time usage and principled decision making.

E1575: **Random coefficient autoregression on trees**
Presenter: **Anand Vidyashankar**, George Mason University, United States

Random coefficient autoregression is frequently used to model time series data exhibiting heterogeneity. In applications arising in biology, finance, and insurance, time series models on a random tree are used to obtain predictions. We formalize the concept of general time series models and provide theoretical results concerning predictions that account for both the randomness in the tree and the correlation between the time series and the tree.

EO356 Room Virtual R30 STATISTICAL INFERENCE OF NETWORK DATA Chair: Jyotishka Datta

E0309: **Fast network community detection with profile-pseudo likelihood methods**
Presenter: **Ji Zhu**, University of Michigan, United States

The stochastic block model is one of the most studied network models for community detection. It is well-known that most algorithms proposed for fitting the stochastic block model likelihood function cannot scale to large-scale networks. One prominent work that overcomes this computational challenge is the one that proposed a fast pseudo-likelihood approach for fitting stochastic block models to large sparse networks. However, this approach does not have a convergence guarantee and is not well suited for small- or medium-scale networks. We propose a novel likelihood-based approach that decouples row and column labels in the likelihood function, which enables a fast alternating maximization; the new method is computationally efficient, performs well for both small and large scale networks, and has a provable convergence guarantee. We show that our method provides strongly consistent estimates of the communities in a stochastic block model. As demonstrated in simulation studies, the proposed method outperforms the pseudo-likelihood approach in terms of both estimation accuracy and computation efficiency, especially for large sparse networks. We further consider extensions of our proposed method to handle networks with degree heterogeneity and bipartite properties.

E0352: **Hypothesis testing and learning on network-valued data**
Presenter: **Debarghya Ghoshdastidar**, Technical University of Munich, Germany

Network analysis has evolved over the past two decades. A traditional view of a network is a tool for modelling interactions among entities of interest; for instance, the analysis of the Facebook network may focus on finding communities of users. Recent applications in bioinformatics and other areas require a perspective where the networks are the quantities of interest. Examples include classification of protein structures as enzyme or non-enzyme, or detecting if brain networks of patients with a neurological disease are statistically different from those of healthy individuals. We refer to such problems as learning from network-valued data to distinguish from the traditional network analysis problems, involving a single

network of interactions. There has been considerable research in supervised learning on network-valued data, with the two most powerful tools being graph kernels and graph neural networks. We focus on two problems beyond the supervised setting: hypothesis testing of large graphs, and clustering network-valued data. A key challenge in such problems is the scarcity of data – typically, one has access to few large graphs, and so popular approaches are not known to work well. We will discuss approaches for network testing and clustering based on ideas from high-dimensional statistics, random graphs and graphons. We will discuss some theoretical properties of these methods (statistical consistency and minimax rates) and demonstrate their empirical performance.

E1022: **Anomalous clique detection via egonets**
Presenter: **Srijan Sengupta**, North Carolina State University, United States

Anomaly in networks often implies illegal or disruptive activity by the actors in the network. Networks can be static, where we have a single snapshot of the system, or dynamic, where we have network snapshots at several points in time. Anomalies can have different meanings in these two scenarios. In static networks, anomaly typically means a local anomaly, in the form of a small anomalous subgraph that is significantly different from the rest of the network. Local anomalies are difficult to detect using simple network-level metrics since the anomalous subnetwork might be too small to cause significant changes to network-level metrics, e.g., network degree. Instead, such anomalies might be detectable if we monitor sub-network level metrics, e.g., degrees of all subgraphs. However, that option is computationally infeasible, as it involves computing total degrees for all $O(2^n)$ subgraphs of an n-node network. We propose a novel anomaly detection method by using egonet p-values, where the egonet of a node is defined as the sub-network spanned by all neighbors of that node. Since there are exactly n egonets, the number of subgraphs being monitored is n, which is a relatively manageable number. We establish the theoretical properties of the egonet method. We demonstrate its accuracy from simulation studies involving a broad range of statistical network models. We also illustrate the method on several well-studied network datasets.

E1064: **Principal component analysis for network samples**
Presenter: **James Wilson**, University of Pittsburgh, United States

The problem of interpretable network representation learning for samples of network-valued data is considered. We propose the Principal Component Analysis for Networks (PCAN) algorithm to identify statistically meaningful low-dimensional representations of a network sample via subgraph count statistics. The PCAN procedure provides an interpretable framework for which one can readily visualize, explore, and formulate predictive models for network samples. We furthermore introduce a fast sampling-based algorithm, sPCAN, which is significantly more computationally efficient than its counterpart, but still enjoys advantages of interpretability. We investigate the relationship between these two methods and analyze their large-sample properties under the common regime where the sample of networks is a collection of kernel-based random graphs. We show that under this regime, the embeddings of the sPCAN method enjoy a central limit theorem and moreover that the population level embeddings of PCAN and sPCAN are equivalent. We assess PCAN's ability to visualize, cluster, and classify observations in network samples arising in nature, including functional connectivity network samples and dynamic networks describing the political co-voting habits of the U.S. Senate. Our analyses reveal that our proposed algorithm provides informative and discriminatory features describing the networks in each sample.

E1070: **Exploratory data analysis for dynamic networks**
Presenter: **Paromita Dubey**, Stanford University, United States

Samples of dynamic or time-varying networks are increasingly encountered in modern data analysis. Common methods for time-varying data such as functional data analysis are infeasible when observations are time courses of networks or other complex non-Euclidean random objects that are elements of general metric spaces. We combat this complexity by a generalized notion of mean trajectory taking values in the object space. For this, we adopt pointwise Frechet means and then construct pointwise distance trajectories between the individual time courses and the estimated Frechet mean trajectory, thus representing the time-varying objects and networks by functional data. Functional principal component analysis of these distance trajectories can reveal interesting features of dynamic networks and object time courses and is useful for downstream analysis. The approach also makes it possible to study the empirical dynamics of time-varying networks, including dynamic regression to the mean or explosive behavior over time. We demonstrate desirable asymptotic properties of sample-based estimators for suitable population targets under mild assumptions. The utility of the proposed methodology is illustrated with Chicago Divvy Bike networks.

EO078 Room Virtual R31 IMAGING DATA ANALYSIS: RECENT DEVELOPMENTS AND APPLICATIONS Chair: Farouk Nathoo

E0993: **Pitfalls in statistical modeling of tumor-infiltrating immune cells from histological imaging**
Presenter: **Finn Hamilton**, BC Cancer, Canada

The advent of successful immunotherapies for cancer has spurred much research to understand the contribution of tumor-infiltrating immune cells (TIL) to patient outcome and to develop new therapies. Despite advances, TIL data arising from histological imaging have multiple issues that are infrequently addressed in modeling studies. These include that TIL are subject to substantial measurement and sampling error, especially when sampled from high-throughput tissue microarrays (TMAs), and that TIL data arise from overdispersed count distributions. Failure to account for these issues can lead to bias in parameter estimation, reduced power, and the failure of findings to generalize to broader populations. We demonstrate how incorporating these considerations into modeling strategies can increase the reliability of inference in TIL studies, and propose key metrics and methods to consider when analyzing TIL or other biomarkers measured from histological imaging.

E1484: **Dimension reduction and regression modelling for imaging genetics**
Presenter: **Farouk Nathoo**, University of Victoria, Canada

Recent advances in technology for brain imaging and high-throughput genotyping have motivated studies examining the influence of genetic variation on brain structure. We describe approaches for the analysis of imaging genetic studies using penalized multi-task regression with priors that provide structured sparsity at both the gene level and SNP level using multivariate Laplace formulations. The model is specified as a three-level Gaussian scale-mixture and we consider both spatial and non-spatial models and Bayesian implementations based on both MCMC and variational Bayes. We also describe an approach for disease-directed dimension reduction of neuroimaging phenotypes based on neural networks. The approaches are evaluated using both simulations as well as test data obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI).

E1487: **Connectivity regression 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, Stewart Mostofsky

Modeling covariance matrices as outcomes has been an important topic in many fields, including in financial and neuroimaging analysis. We consider the problem of regressing covariance matrices on vector covariates, collected from each observational unit. The main aim is to uncover the variation in the covariance matrices across units that are explained by the covariates. A covariate-assisted principal regression framework is introduced that identifies the components predicted by the covariates through a generalized linear model type link function. We develop computationally efficient optimization algorithms to jointly search the linear projections of the covariance matrices as well as the regression coefficients, and we establish the asymptotic properties. Using extensive simulation studies, our method shows higher accuracy and robustness in coefficient estimation than competing methods. We will illustrate extensions of this framework for high dimensional and longitudinal settings. Applied to a resting-state functional magnetic resonance imaging study, our approach identifies the human brain network changes associated with covariates.

E0434: **A knowledge-based multivariate method for examining gene-brain-behavioural/cognitive relationships**
Presenter: **Heungsun Hwang**, McGill University, Canada

With advances in neuroimaging and genetics, imaging genetics is a naturally emerging field that combines genetic and neuroimaging data with behavioural or cognitive outcomes to examine the genetic influence on altered brain functions associated with behavioural or cognitive variation. We propose a statistical approach, termed imaging genetics generalized structured component analysis (IG-GSCA), which allows researchers to investigate such gene-brain-behaviour/cognitive associations, taking into account well-documented biological characteristics (e.g., genetic pathways, gene-environment interactions, etc.) and methodological complexities (e.g., multicollinearity) in imaging genetic studies. We describe the conceptual and technical underpinnings of IG-GSCA and provide its application for investigating how several depression-related genes and their interactions with an environmental variable (experience of potentially traumatic events) might influence the thickness variations of 53 brain regions, which in turn tended to affect depression severity in a sample of Korean participants.

EO810 Room Virtual R35 STATISTICAL INFERENCE FOR COMPLEX DATA Chair: Anderson Ye Zhang

E0379: **BEAUTY powered BEAST**

Presenter: **Kai Zhang**, University of North Carolina at Chapel Hill, United States

Co-authors: Zhigen Zhao, Wen Zhou

Nonparametric dependence detection is studied with the proposed binary expansion approximation of uniformity (BEAUTY) approach, which extends the celebrated Euler's formula, and approximates the characteristic function of any copula distribution with a linear combination of means of binary interactions from marginal binary expansions. This novel theory enables the unification of many important existing tests through an approximation from some quadratic form of symmetry statistics, where the deterministic weight matrix characterizes the power properties of each test. To achieve a robust high power, we study test statistics with data-adaptive weights, referred to as the binary expansion adaptive symmetry test (BEAST). By utilizing the properties of the binary expansion filtration, we show that the Neyman-Pearson test of uniformity can be approximated by an oracle weighted sum of symmetry statistics. The BEAST with this oracle leads all existing tests we considered in empirical power against all forms of alternatives, thus sheds light on the potential of substantial improvements in power and on the form of optimal weights under each alternative. To approach this oracle power, we develop the BEAST through a regularized subsampling approximation of the oracle test. The BEAST improves the empirical power of many existing tests against a wide spectrum of common alternatives while providing a clear interpretation of the form of dependency upon rejection.

E0433: **Statistical inference of robust regression with contaminated errors**

Presenter: **Zhao Ren**, University of Pittsburgh, United States

Co-authors: Wenxin Zhou, Peiliang Zhang

The robust estimation and inference problems for linear regression are studied in the increasing dimension regime. Given a random design, we consider the conditional distributions of error terms are contaminated by some arbitrary distribution (possibly depending on the covariates) with proportion ε but otherwise can also be heavy-tailed and asymmetric. We show that simple robust M-estimators such as Huber and smoothed Huber, with an additional intercept added in the model, can achieve the minimax rates of convergence under the l_2 loss. In addition, two types of confidence intervals with root-n consistency are provided by a multiplier bootstrap technique when the necessary condition on contamination proportion $\varepsilon = o(1/\sqrt{n})$ holds. For a larger ε, we further propose a debiasing procedure to reduce the potential bias caused by contamination, and prove the validity of the debiased confidence interval. At last, we extend our methods to the communication-efficient distributed estimation and inference setting. A comprehensive simulation study exhibits the effectiveness of our proposed inference procedures.

E0591: **Self-supervised metric learning in multi-view data: A downstream task perspective**

Presenter: **Shulei Wang**, University of Illinois at Urbana-Champaign, United States

Self-supervised metric learning has been a successful approach for learning a distance from an unlabeled dataset. The resulting distance is broadly useful for improving various distance-based downstream tasks, even when no information from downstream tasks is utilized in the metric learning stage. To gain insights into this approach, we develop a statistical framework to theoretically study how self-supervised metric learning can benefit downstream tasks in the context of multi-view data. Under this framework, we show that the target distance of metric learning satisfies several desired properties for the downstream tasks. On the other hand, our investigation suggests the target distance can be further improved by moderating each direction's weights. In addition, our analysis precisely characterizes the improvement by self-supervised metric learning on four commonly used downstream tasks: sample identification, two-sample testing, k-means clustering, and k-nearest neighbor classification.

E0555: **Contiguity under high-dimensional Gaussianity with applications to covariance testing**

Presenter: **Yandi Shen**, University of Chicago, United States

Co-authors: Qiyang Han, Tiefeng Jiang

Le Cams third/contiguity lemma is a fundamental probabilistic tool in mathematical statistics for several major developments in classical statistical estimation and testing theories. Despite widespread applications to low-dimensional statistical problems, the stringent requirement of Le Cams third/contiguity lemma on the asymptotic distributional expansions of the log-likelihood makes it challenging to use in many modern high-dimensional statistical problems. A non-asymptotic analogue of Le Cams third/contiguity lemma is established under high dimensional normal populations, which requires only mean and variance bounds of the statistic under study, but without an exact distributional expansion of the likelihood ratio. As a demonstration of the power of the new contiguity result, we obtain asymptotically exact power formulae for a number of widely used high-dimensional covariance tests, including the likelihood ratio tests and trace tests, that hold uniformly over all possible alternatives under mild growth conditions on the dimension-to-sample ratio. These new results go far beyond the scope of previous available case-specific techniques, and exhibit new phenomena regarding the behavior of these important class of covariance tests.

EO736 Room Virtual R37 ADVANCES IN MULTIVARIATE METHODS Chair: Mengxi Yi

E0520: **On robust estimates of sphericity in high-dimension**

Presenter: **Esa Ollila**, Aalto University, Finland

Co-authors: Elias Raninen

The need to estimate or test sphericity (i.e., that the covariance matrix is proportional to identity) arises in various applications in statistics, and thus the problem has been investigated in numerous papers, most recently in shrinkage covariance matrix estimation problems. We investigate robust estimates of sphericity, especially in a high-dimensional setting, where the data dimensionality p is larger or of similar magnitude as the sample size n. The population measure of sphericity that we consider is defined as the ratio of the mean of the squared eigenvalues of the covariance matrix relative to the mean of its eigenvalues squared. This population quantity is then estimated using robust spatial (both symmetrized and standard) sign and rank covariance matrices as well as using robust M-estimators of scatter matrices. Properties of such estimators are derived and reviewed, and a simulation study is provided comparing the robust estimators of sphericity in the high-dimensional setting.

E0521: **Matrix completion with model-free weighting**

Presenter: **Xiaojun Mao**, Fudan University, China

A novel method is proposed for matrix completion under general non-uniform missing structures. By controlling an upper bound of a novel balancing error, we construct weights that can actively adjust for the non-uniformity in the empirical risk without explicitly modeling the observation probabilities, and can be computed efficiently via convex optimization. The recovered matrix based on the proposed weighted empirical risk enjoys

appealing theoretical guarantees. In particular, the proposed method achieves stronger guarantee than existing work in terms of the scaling with respect to the observation probabilities, under asymptotically heterogeneous missing settings (where entry-wise observation probabilities can be of different orders). These settings can be regarded as a better theoretical model of missing patterns with highly varying probabilities. We also provide a new minimax lower bound under a class of heterogeneous settings. Numerical experiments are also provided to demonstrate the effectiveness of the proposed method.

E0625: **High quantile regression for tail dependent time series**
Presenter: **Ting Zhang**, University of Georgia, United States

Quantile regression serves as a popular and powerful approach for studying the effect of regressors on quantiles of a response distribution. However, existing results on quantile regression were mainly developed when the quantile level is fixed, and the data are often assumed to be independent. Motivated by recent applications, we consider the situation where (i) the quantile level is not fixed and can grow with the sample size to capture the tail phenomena; and (ii) the data are no longer independent but collected as a time series that can exhibit serial dependence in both tail and non-tail regions. To study the asymptotic theory for high quantile regression estimators in the time series setting, we introduce a previously undescribed tail adversarial stability condition, and show that it leads to an interpretable and convenient framework for obtaining limit theorems for time series that exhibit serial dependence in the tail region but are not necessarily strong mixing. Numerical experiments are provided to illustrate the effect of tail dependence on high quantile regression estimators, where simply ignoring the tail dependence may lead to misleading p-values.

E1295: **Best subset selection is robust against design dependence**
Presenter: **Ziwei Zhu**, University of Michigan, Ann Arbor, China
Co-authors: Jianqing Fan, Yongyi Guo

Best subset selection (BSS) is widely known as the holy grail for high-dimensional variable selection. We investigate the variable selection properties of BSS when its target sparsity is greater than or equal to the true sparsity. The main message is that BSS is robust against design dependence in terms of achieving model consistency and sure screening, and more importantly, that such robustness can be propagated to the near best subsets that are computationally tangible. Specifically, we introduce an identifiability margin condition that is free of restricted eigenvalues and show that it is sufficient and nearly necessary for BSS to exactly recover the true model. A relaxed version of this condition is also sufficient for BSS to achieve the sure screening property. Moreover, we show that a two-stage fully corrective iterative hard thresholding (IHT) algorithm can provably find a near best subset within logarithmic steps; another round of exact BSS within this set can recover the true model. The simulation studies and real data examples show that IHT yields lower false discovery rates and higher true positive rates than the competing approaches including LASSO, SCAD and Sure Independence Screening (SIS), especially under highly correlated design.

E1680: **Large precision matrix estimation for compositional data**
Presenter: **Shucong Zhang**, University of International Business and Economics, China

High dimensional compositional data are prevalent in many applications. The simplex constraint poses intrinsic challenges to inferring the conditional dependence relationships among the components forming a composition, as encoded by a large precision matrix. We introduce a precise specification of the compositional precision matrix and relate it to its basis counterpart, which is shown to be asymptotically identifiable under suitable sparsity assumptions. By exploiting this connection, we propose a composition adaptive regularized estimation (CARE) method for estimating the sparse basis precision matrix. We derive rates of convergence for the estimator and provide theoretical guarantees on support recovery and data-driven parameter tuning. Our theory reveals an intriguing trade-off between identification and estimation and highlights the blessing of dimensionality for compositional data. In particular, in sufficiently high dimensions, the CARE estimator achieves minimax optimality and performs as well as if the basis were observed. The advantages of CARE over existing methods are illustrated by simulation studies and an application to inferring microbial ecological networks in the human gut.

EO651 Room Virtual R39 RECENT ADVANCES ON STOCHASTIC MODELING Chair: Pepa Ramirez Cobo

E1197: **Modeling with the MAP counting process**
Presenter: **Rosa Lillo**, Universidad Carlos III de Madrid, Spain
Co-authors: Pepa Ramirez Cobo, Marcos Gonzalez

It is known that Markovian arrival processes (MAPs) are very suitable processes for stochastic modeling, among other things because they allow dependent inter-arrival times. This property appears, for example, in the data relating to modern call centers, which are characterized by non-negligible dependence patterns and by significant changes in arrival rates throughout the day. Most of the previous statistical approaches for MAPs are based on the distribution of the inter-arrival times, but in many cases, it is more interesting and practical to take the counting process into account. For this reason, the inference of MAPs processes is approached from the perspective of the associated counting process, of which almost everything is unknown except for the closed-form expression of the counting process' covariance function. New properties concerning the correlation patterns and monotonicity shall be illustrated.

E1498: **Bayesian modelling of the selection bias problem in regression**
Presenter: **M Remedios Sillero-Denamiel**, Trinity College Dublin, Ireland
Co-authors: Simon Wilson, Hieu Cao

In the regression setting, it is typically assumed that training and test sets follow similar distributions, but that is not always true, as is the case with the sky surveys of galaxies where faint ones are not observed in favour of brighter ones. In addition, when data follow complicated non-Gaussian distributions, the full conditional density has to be estimated to properly quantify the uncertainty in the predictions. We present a Bayesian approach to estimate the conditional density under selection bias.

E1709: **Modeling a repairable discrete multi-state system with a vacation policy**
Presenter: **Juan Eloy Ruiz-Castro**, University of Granada, Spain
Co-authors: Mohammed Dawabsha

A complex multi-state system that evolves in discrete-time and is subject to multiple events and preventive maintenance is considered. Various internal levels of degradation are assumed. The repair facility is composed of a repairperson, who may take one or more vacations during the period considered. A vacation policy is established for the repairpersons vacation time. Two different tasks may be performed by the repairperson: corrective repair and preventive maintenance. Phase type distributions are considered in the modelling. The transient and stationary distributions are determined and several reliability measures are developed in a matrix-algorithmic form. Rewards and costs are included in the model.

E1714: **Bivariate exponential distribution with atoms at zero**
Presenter: **Clara Gardner**, Technical University of Denmark, Denmark
Co-authors: Bo Friis Nielsen

Raftery's bivariate exponential distribution, the Farlie-Gumbel-Morgenstern construction and the Kibble-distribution are three different models for bivariate exponential distributions, which all have a multivariate PH distribution of the MPH* type. The models are expanded to include atoms at 0 such that they can model non-negative data - e.g. waiting times. An advantage is that these modified models also will have an MPH* representation. For these three models, two questions are investigated: 1) How to perform parameter estimation? 2) Which types of data are most suited for each model? For the first question two estimation techniques are investigated, namely raw maximum likelihood estimation and the EM-algorithm for

MPH* models. The two methods are compared both in terms of precision and computational time. The second question is not as easy to answer, and two different approaches are taken. Firstly, the stability of the models in the parameter space is investigated. Both the estimation errors and the variation in terms of the curvature are considered. Secondly, the versatility of the models is checked by fitting data stemming from one model to another model. At last, as a real-life application, the three models are tested on data of delays in public transport.

E1687: **Latent variable modelling in the number of kinds problem**
Presenter: **Simon Wilson**, Trinity College Dublin, Ireland
Co-authors: Asmaa Al-Ghamdi

Data in a number of problems are usually modelled as one of 2 types: complete sampling data, where the number of individuals sampled and their kind are observed, and temporal data that describe when new kinds were observed but lack information on numbers sampled. Very different models and estimation methods apply to these types. Inconveniently, one of the most important applications of this problem, estimation of the number of species, falls into neither type; complete sampling information is lacking, but there is some proxy information on it (typically some measure of effort like estimates of numbers of individuals that can be sampled). We propose a hybrid model that allows such proxy information to be incorporated. The advantage of this approach is that it produces a framework around which the uncertainties in the number of species estimation can be modelled and quantified, something that is certainly needed for a question where estimates vary by at least an order of magnitude and estimates of uncertainty are often lacking. The inference is implemented via ABC and applied to 2 large databases: Catalogue of Life and World Register of Marine Species. Prior sensitivity and approaches to speeding up the implementation are discussed.

EO076 Room Virtual R40 COMPOSITIONAL, DISTRIBUTIONAL AND RELATIVE ABUNDANCE DATA I Chair: Karel Hron

E0955: **Classification of compositional data with selective pivot coordinates**
Presenter: **Karel Hron**, Palacky University, Czech Republic
Co-authors: Julie Rendlova, Peter Filzmoser

In classification tasks with geochemical of chemometric data, it frequently happens that observations are of relative (compositional) nature. The logratio approach to compositional data analysis offers a concise methodology by replacing the original scale-invariant positive data by reasonable real variables. The preferred type of such logratio variables corresponds to orthonormal coordinates where the first coordinate aggregates all logratios with the specific part of interest and can be thus linked to that component - we refer to so-called pivot coordinates. However, including all respective logratios into the first pivot coordinate may lead to an artificial occurrence of false positives in biomarker detection. Therefore, we propose a method excluding aberrant logratios so that the coordinate which is afterwards considered to be the pivot one in the resulting coordinate system contains already just the "cleaned" information about the relative dominance of the specific component. Importantly, the alternative choice of pivot coordinates, which we suggest to call selective pivot coordinates, does not influence the quality of classification itself since both coordinate systems are just rotations of each other. The effect of such a choice of coordinates will be presented with the partial least squares regression - discriminant analysis of metabolomic data.

E1010: **The influence of the error-propagation on ilr-base selection for compositional models**
Presenter: **Joseph Sanchez-Balseca**, Universitat Politecnica de Catalunya, Spain
Co-authors: Agusti Perez-Foguet

Most data in the geoenvironmental sciences are compositional. If this data is not treated adequately, the results obtained from its numerical modelling could be wrong. The log-ratio transformation was proposed to resolve the principal statistical problems related to compositional data (CoDa). The isometric-log-ratio transformation (ilr) is the most used due to its advantage to represent the simplex space orthogonally. For this purpose, it is required to define an orthonormal base, and one method is through the Sequential Binary Partition (SBP) that is an iterative process. However, recent modelling researches with a compositional approach applied to energy and air pollution has found that different select order in the SBP produces different results. This behaviour from the practical numerical modelling could contradict the theory of CoDa. The aim is to explain the problem through an error-propagation approach. It is necessary to consider the finite arithmetic implicated in the numeric modelling software. The results are exposed using air pollution data.

E0962: **Identification of significant pairwise logratios in compositional data based on sparse PCA**
Presenter: **Viktorie Nesrstova**, Palacky University, Olomouc, Czech Republic
Co-authors: Karel Hron, Peter Filzmoser, Ines Wilms

Compositional data consist of observations that carry relative information in the ratios between the parts. They are commonly expressed as log-ratio coordinates with respect to an orthonormal basis of their sample space. However, for modern high-dimensional data, analyzing all pairwise logratios rapidly becomes infeasible due to their sheer size. The focus is on the high-dimensional analysis by identifying the most significant pairwise logratios. To this end, sparse PCA is leveraged when constructing backward pivot coordinates that highlight a pairwise logratio within a complete set of logratio coordinates. The significant pairwise logratios are identified while balancing sufficient sparsity of the resulting loadings and explained variance. The performance of the procedure is demonstrated both in simulation and applications on real-world data.

E0533: **Too big to fail? An analysis of the Colombian banking system through compositional data**
Presenter: **Juan Vega Baquero**, Universitat de Barcelona, Spain
Co-authors: Miguel Santolino

Although still incipient in economics and finance, compositional data analysis (in which relative information is more important than absolute values) has become more relevant in statistical analysis in recent years. A concentration index for financial/banking systems is constructed by means of compositional analysis, to establish the potential existence of "too big to fail" financial entities. The intention is to provide an early warning tool for regulators about this kind of institution. The index has been applied to the Colombian banking system and assessed over time with a forecast to determine whether the system is becoming more concentrated or not. It was found that the concentration index has been decreasing in recent years and the model predicts that this trend will continue. In terms of the methodology used, compositional models were shown to be more stable and to lead to better prediction of the index than the classical multivariate methodologies.

E1465: **Regression analysis using compositional balances: Case study for chronic kidney disease and environmental toxins**
Presenter: **Jennifer McKinley**, Queen's University Belfast, United Kingdom
Co-authors: Ute Mueller, Pete Atkinson, Damian Fogarty

Investigating the importance of multi-element interactions of environmental toxins in understanding the occurrence of clusters of chronic diseases, such as chronic kidney disease, is of global concern. Environmental toxins (air, soil and waterborne) often comprise soil or water geochemical data, which are compositional in nature in that they convey relative information that should be extracted by treating log-ratio or equivalently transformed data. For regression analysis involving compositional, the concept of balances between two groups of parts of a composition provides an interpretable approach to identify components whose relative abundances may be associated with a response variable. Several approaches to select balances are available, which constitute data- or knowledge-driven approaches within a compositionally-compliant context. However, many questions remain on how to interpret compositional balances including the impact of the ordering of elements in a compositional balance within a data-driven or knowledge informed knowledge-driven approach. Moreover, regression models assume independence between the observations, an assumption that may not be valid for spatial data. The aim is to explore different compositional balance approaches and the impact of spatial

dependence, using a case study on chronic kidney disease and its relation with air, soil and waterborne environmental toxins.

EO702 Room K2.40 (Hybrid 08) ADVANCES IN EMPIRICAL BAYES METHODOLOGY (VIRTUAL) Chair: Asaf Weinstein

E0687: **Invidious comparisons: Ranking and selection as compound decisions**
Presenter: **Jiaying Gu**, University of Toronto, Canada
Co-authors: Roger Koenker

There is an innate human tendency, one might call it the league table mentality, to construct rankings. Schools, hospitals, sports teams, movies, and myriad other objects are ranked even though their inherent multi-dimensionality would suggest that - at best - only partial orderings were possible. We consider a large class of elementary ranking problems in which we observe noisy, scalar measurements of merit for n objects of potentially heterogeneous precision and are asked to select a group of the objects that are most meritorious. The problem is naturally formulated in the compound decision framework empirical Bayes theory, but it also exhibits close connections to the recent literature on multiple testing. The nonparametric maximum likelihood estimator for mixture models is employed to construct optimal ranking and selection rules. Performance of the rules is evaluated in simulations and an application to ranking U.S kidney dialysis centers.

E1705: **Confidence intervals for nonparametric empirical Bayes analysis**
Presenter: **Nikolaos Ignatiadis**, Stanford University, United States
Co-authors: Stefan Wager

In an empirical Bayes analysis, we use data from repeated sampling to imitate inferences made by an oracle Bayesian with extensive knowledge of the data-generating distribution. Existing results provide a comprehensive characterization of when and why empirical Bayes point estimates accurately recover oracle Bayes behavior. We develop flexible and practical confidence intervals that provide asymptotic frequentist coverage of empirical Bayes estimands, such as the posterior mean or the local false sign rate. The coverage statements hold even when the estimands are only partially identified or when empirical Bayes point estimates converge very slowly.

E1744: **Transfer learning for empirical bayes estimation: A nonparametric integrative Tweedie approach**
Presenter: **Wenguang Sun**, University of Southern California, United States

Compound estimation of normal means with auxiliary data collected from related source domains is considered. The empirical Bayes framework provides an elegant interface to pool information across different samples and construct efficient shrinkage estimators. We propose a nonparametric integrative Tweedie (NIT) approach to transferring structural knowledge encoded in the auxiliary data from related source domains to assist the simultaneous estimation of multiple parameters in the target domain. Our transfer learning algorithm uses convex optimization tools to directly estimate the gradient of the log-density through an embedding in the reproducing kernel Hilbert space (RKHS), which is induced by the Steins discrepancy metric. Most popular structural constraints can be easily incorporated into our estimation framework. We characterize the asymptotic L_p risk of NIT by first rigorously analyzing its connections to the RKHS risk, and second establishing the rate at which NIT converges to the oracle estimator. The improvements in the estimation risk and the deteriorations in the learning rate are precisely tabulated as the dimension of side information increases. The numerical performance of NIT and its superiority over existing methods are illustrated through the analysis of both simulated and real data.

E1759: **On permutation invariant problems in large-scale inference**
Presenter: **Asaf Weinstein**, Hebrew University of Jerusalem, Israel

A class of simultaneous inference problems is considered that are invariant under permutations, meaning that all components of the problem are oblivious to the labelling of the multiple instances under consideration. For any such problem we identify the optimal solution which is itself permutation invariant, the most natural condition one could impose on the set of candidate solutions. Interpreted differently, for any possible value of the parameter we find a tight (non-asymptotic) lower bound on the statistical performance of any procedure that obeys the aforementioned condition. By generalizing the standard decision-theoretic notions of permutation invariance, we show that the results apply to a myriad of popular problems in simultaneous inference, so that the ultimate benchmark for each of these problems is identified. The connection to the nonparametric empirical Bayes approach of Robbins is discussed in the context of asymptotic attainability of the bound uniformly in the parameter value.

CO824 Room K0.16 (Hybrid 02) MACHINE LEARNING FOR FINANCE: THEORY AND APPLICATION Chair: Haoyang Cao

C0236: **Optimal execution of foreign securities: a double-execution problem with signatures and machine learning**
Presenter: **Leandro Sanchez-Betancourt**, University of Oxford, United Kingdom
Co-authors: Alvaro Cartea

The expected signature of equity and foreign exchange markets is employed to derive an optimal double-execution trading strategy. The signature of a path of a stochastic process is a sequence of real numbers that provides a full description of the evolution of the process. The double-execution strategy maximises the wealth (in units of the domestic currency) of an investor who liquidates a block of shares in a foreign stock market. Our approach is model agnostic because we do not specify the dynamics of the market. We prove that the optimal strategy is a linear combination of the terms in the expected signature of the market and employ high-frequency data from Nasdaq and for various currencies to compute the signature of the market. Data for ten stocks and four currency pairs are employed to implement the strategy. Our results show that the performance of the signature-based double-execution strategy is superior to the performance of the benchmarks. In most cases, the outperformance increases further when the signature of the market is enhanced with the price dynamics of the SPY - a tracker of the Standard and Poor's 500 index.

C0318: **Policy gradient methods find the Nash equilibrium in N-player general-sum linear-quadratic games**
Presenter: **Huining Yang**, University of Oxford, United Kingdom
Co-authors: Ben Hambly, Renyuan Xu

Policy optimization algorithms have achieved substantial empirical successes in addressing a variety of non-cooperative multi-agent problems, including self-driving vehicles, real-time bidding games, and optimal execution in financial markets. However, there have been few results from a theoretical perspective showing why such a class of reinforcement learning algorithms performs well with the presence of competition among agents. We explore the natural policy gradient method for a class of N-agent general-sum linear-quadratic games. We provide a global linear convergence guarantee for this approach in the setting of finite time horizon and stochastic dynamics when there is a certain level of noise in the system. The noise can either come from the underlying dynamics or carefully designed explorations from the agents. We illustrate our results with numerical experiments to show that even in situations where the policy gradient method may not converge in the deterministic setting, the addition of noise leads to convergence.

C0321: **Interactions of market making algorithms: A study on perceived collusion**
Presenter: **Wei Xiong**, University of Oxford, United Kingdom

The widespread use of market-making algorithms and the associated feedback effects may have unexpected consequences which need to be better understood. In particular, the phenomenon of 'tacit collusion' in which the interaction of algorithms leads to an outcome similar to collusion among market makers, has increasingly received regulatory scrutiny. We propose a game-theoretic model of a financial market in which multiple market-makers compete for market share and learn from market data to adjust their spreads. We model this learning process through a decentralized

multi-agent reinforcement learning algorithm and show that, even in absence of information sharing, market prices may converge to levels that are similar to a collision situation, resulting in 'tacit collusion'. We briefly discuss the implications of our research for market regulators.

C0323: **Analysis and modeling of client order flow in limit order markets**
Presenter: **Felix Prenzel**, University of Oxford, United Kingdom

Orders in major electronic stock markets are organised through centralised limit order books (LOBs). Large amounts of historical data have led to extensive research modelling LOBs to understand their dynamics better and build simulators as a framework for controlled experiments, when testing trading algorithms or execution strategies. Most work in the literature models the aggregate view of the LOB, also known as queue size, using a point process. Brokers and exchanges, however, also have more granular information on the origin of limit orders. This leads to a more granular view of limit order book dynamics, which we attempt to model using a heterogeneous model of order flow. We present a granular representation of the limit order book that allows accounting for the origins of limit orders. Using trade execution data, we analyse the properties of variables in this representation. The heterogeneity of order flow is modelled by segmenting traders into different clusters, for which we identify representative prototypes. This segmentation appears to be stable both over time as well as over different stocks. Our findings can be leveraged to build more realistic order flow models that account for the diversity of the market participants.

C0552: **Identifiability in inverse reinforcement learning**
Presenter: **Haoyang Cao**, The Alan Turing Institute, United Kingdom

Inverse reinforcement learning attempts to reconstruct the reward function in a Markov decision problem, using observations of agent actions. As already observed in earlier works the problem is ill-posed, and the reward function is not identifiable, even under the presence of perfect information about optimal behavior. We provide a resolution to this non-identifiability for problems with entropy regularization. For a given environment, we fully characterize the reward functions leading to a given policy and demonstrate that, given demonstrations of actions for the same reward under two distinct discount factors, or under sufficiently different environments, the unobserved reward can be recovered up to a constant. Through a simple numerical experiment, we demonstrate the accurate reconstruction of the reward function through our proposed resolution.

CO044 Room Virtual R18 DYNAMIC MODELS WITH REGIME SWITCHING Chair: Willi Semmler

C1167: **Instability in regime switching models**
Presenter: **Pu Chen**, Melbourne Institute of Technology, Australia
Co-authors: Chihying Hsiao, Willi Semmler

The purpose is to investigate the instability in a self-exciting regime-switching autoregressive model, in particular, those regime-switching models that are locally stable in each of their regimes respectively. It turns out that the stability locally in each of the regimes is not sufficient to guarantee the stability of the model. The mechanism of the instability is explained and a sufficient condition for the instability is provided.

C1467: **Pandemic meltdown and economic recovery: A multi-phase dynamic model, empirics, and policy**
Presenter: **Willi Semmler**, New School for Social Research, United States

A two-phase model of a Pandemic meltdown is studied. In the first phase, the spread of Pandemic disease and the effects of a lockdown policy are explored. The output gap either starting from a positive gap (boom period), or negative gap (recession period), does not converge toward the potential output, with the output gap zero, but will stay below it for a considerable time period. This arises from the nonlinearity effects of the lockdown decisions on the output gap, infection and fatality rates. Yet, given a still large fraction of susceptible population in a second phase, the efficacy of an expansionary monetary policy is studied. We explore to what extent and speed it can come to the rescue and help to move the economy out of the meltdown, without significant fatalities. We thus suggest a multi-phase macro model with phase shift where in the first phase the diffusion of the infectious disease, lockdown, fatalities and output decline is dominant. In the second phase of a policy-induced economic recovery, there is still a dynamic interaction of the output gap with the spread of the Pandemic disease though to a lesser extent. The two phases are studied in a regime change model where the state variables and objective functions are allowed to change from the first stage to the second one. The two-phase finite horizon decision model is empirically calibrated and numerically solved through AMPL, a new solution method for finite-horizon dynamic models.

C1181: **Recession-specific recoveries: L's, U's and everything in between**
Presenter: **Irina Panovska**, University of Texas at Dallas, United States
Co-authors: Luiggi Donayre

The assumption that recessions are all alike is relaxed and a new model of output growth is proposed that allows for recession-specific recoveries. Output growth is modelled as the weighted average of Markov-switching processes that temporarily alter the level of real GDP (U-shaped) and those with permanent effects (L-shaped), where the recession-specific weight is endogenously estimated. Only the 1969-70 and 2007-09 recessions are characterized exclusively as U and L, respectively. The other 85% of U.S. recessions reflect a weighted combination of the two shapes. Consequently, models that imply only one possible path for a given recession may be insufficient to fully characterize the behavior of output during recessionary periods. With respect to fitting output growth, our model outperforms those that generate either U- or L-shaped recoveries and the model-implied paths closely track the level of actual U.S. real GDP during recessions and recoveries.

C1192: **Hamilton versus Hamilton: Spurious nonlinearities**
Presenter: **Luiggi Donayre**, University of Minnesota - Duluth, United States

Using Monte Carlo simulations, the purpose is to evaluate the ability of the Hamilton Decomposition (HD) approach into trend and cycle to adequately identify asymmetries in business cycles fluctuations. By considering different specifications of linear and asymmetric processes consistent with previous estimates, the results indicate that the HD approach is unable to preserve true asymmetric behavior nor reproduce U.S. business cycles features, especially in highly persistent or mildly asymmetric processes, or in small samples. The findings are robust to the presence of a time-varying drift, the complexity of the autoregressive dynamics and symmetric nonlinearity. Furthermore, the HD approach generates spurious expansionary periods when none exist in the data-generating process. Interestingly, they occur, exclusively, in the case of Markov-switching models, but not for other nonlinear models. Meanwhile, the distortions are also present in the case of symmetric nonlinearity. Based on these findings, caution is called into question when the approach is applied to processes that are thought to behave nonlinearly.

C1150: **The effects of money-financed fiscal stimulus in a small open economy**
Presenter: **Eiji Okano**, Nagoya City University, Japan
Co-authors: Masataka Eguchi

The purpose is to analyze the effects of money-financed (MF) fiscal stimulus and compare them with those resulting from a conventional debt-financed (DF) fiscal stimulus in a small open economy. We find that in normal times, MF fiscal stimulus is effective in increasing output. In a liquidity trap where the ZLB is applicable, even though the decrease in both consumer price index (CPI) inflation and output is more severe than in a closed economy when there is no fiscal response, MF fiscal stimulus is effective in stabilizing both. Accordingly, we show that even in an imperfect pass-through environment including a liquidity trap, an increase in government expenditure under MF fiscal stimulus is effective. In contrast, our policy implications concerning an increase in government expenditure under DF fiscal stimulus lie opposite to previous work assuming a closed economy. In normal times, an increase in government expenditure under the DF scheme in a small open economy is more effective than in a closed economy, although it has been argued that it is much less effective. In a liquidity trap, an increase in government expenditure under the DF

scheme is less effective, also in contrast to previous work. We find that even in an imperfect pass-through environment, an increase in government expenditure under DF fiscal stimulus is not effective.

CO186 Room Virtual R32 ADVANCES IN EMPIRICAL MACROECONOMICS Chair: Christian Matthes

C0437: **A neural Phillips curve**
Presenter: **Philippe Goulet Coulombe**, University of Pennsylvania, United States
Many problems plague the estimation of the New Keynesian Phillips curve. Amongst them is the hurdle that the two key components, inflation expectations and the output gap, are both unobserved. Traditional remedies include creating reasonable proxies for the notable absentees or extracting them via some form of assumptions-heavy filtering procedure. Essentially, this is all unsupervised learning. We move towards a supervised extraction of inflation key drivers by developing a Hemisphere Neural Network whose peculiar structure allows the interpretation of the last layer's cells output as key macroeconomic latent states. Many benefits come from this neural approach. First, nonlinearities are trivially allowed for. Second, computations are quick and done within standard deep learning software. Lastly, the model typically forecast better than a wide array of benchmarks (including plain neural nets) while being interpretable.

C0543: **Averaging impulse responses**
Presenter: **Christian Matthes**, Indiana University, United States
Co-authors: Paul Ho, Thomas Lubik
Impulse response analysis is a key tool for researchers to study the effects of shocks and policies on economic outcomes. When estimating impulse responses, economists have a wide range of options. For example, one can choose between local projections (LP) and vector autoregressions (VARs), Bayesian and frequentist methods, and different specifications. Each choice has its own drawbacks and benefits. It is well known that these choices can generate significantly different results. While there is a growing literature discussing conditions for which one approach might be preferred over another, in practical applications many of the conditions are likely to be difficult to verify. We propose the use of prediction pools to average impulse responses across different models.

C0579: **Time Varying IV-SVARs and the effects of monetary policy on financial variables**
Presenter: **Robin Braun**, Bank of England and London School of Economics, United Kingdom
Co-authors: George Kapetanios, M. Marcellino
Parameter instability is pervasive in economics but there is no consensus on the best way to model it, as that depends on the specific process driving parameter changes. We develop kernel-based estimation for time-varying vector autoregressive models identified by instrumental variables (SVAR-IV), a flexible method that allows general (non-parametric) parameter evolution. Specifically, we derive the asymptotic distribution for parameter estimates and impulse responses of various SVAR-IV representations. We then apply the method to study the changes in the effects of monetary policy on financial variables.

C0869: **Bayesian estimation of epidemiological models: Methods, causality, and policy trade-offs**
Presenter: **Jonas Arias**, Federal Reserve Bank of Philadelphia, United States
Co-authors: Jesus Fernandez-Villaverde, Juan Rubio-Ramirez, Minchul Shin
A general framework is presented for Bayesian estimation and causality assessment in epidemiological models. The key is the use of sequential Monte Carlo methods to evaluate the likelihood of a generic epidemiological model. Once we have the likelihood, we specify priors and rely on a Markov chain Monte Carlo to sample from the posterior distribution. We show how to use the posterior simulation outputs as inputs for exercises in causality assessment. We apply the approach to Belgian data for the COVID-19 epidemic during 2020. The estimated time-varying-parameters SIRD model captures the data dynamics very well, including the three waves of infections. We use the estimated (true) number of new cases and the time-varying effective reproduction number from the epidemiological model as information for structural vector autoregressions and local projections. We document how additional government-mandated mobility curtailments would have reduced deaths at zero cost or a very small cost in terms of output.

C0994: **A temporary VAT cut as unconventional fiscal policy: Evidence from Germany**
Presenter: **Benjamin Born**, Frankfurt School of Finance & Management, Germany
Co-authors: Ruediger Bachmann, Olga Goldfayn-Frank, Georgi Kocharkov, Ralph Luetticke, Michael Weber
The impact of the temporary VAT cut, meant to stimulate the Covid-19 stricken German economy, during the second half of 2020 is evaluated. Survey data help us to overcome the identification problem by allowing us to classify households according to their subjective perception of this policy measure. We find that the temporary VAT cut led to a substantial relative increase in durable spending. For example, households with a high perceived pass-through into consumer prices spent about 40% more than those with low or no perceived pass-through. Using scanner data, we also find that semi- and non-durable spending increased.

CO248 Room Virtual R33 ADVANCES IN MACROECONOMETRICS Chair: Mikkel Plagborg-Moller

C0363: **Testing macroeconomic policies with sufficient statistics**
Presenter: **Geert Mesters**, Universitat Pompeu Fabra, Spain
Co-authors: Regis Barnichon
The evaluation of macroeconomic policy decisions has traditionally relied on the formulation of a specific economic model. We present a framework to assess policy decisions with minimal assumptions on the underlying structure of the economy. Given a policy maker's loss function, we propose a statistic —the Optimal Policy Perturbation (OPP)— to test (i) whether a policy decision is optimal, i.e., whether it minimizes the loss function, and (ii) the optimality of the policy maker's reaction function, i.e., the optimality of the systematic conduct of policy over some period of time. The computation of the OPP does not rely on specifying an underlying model, and it can be computed from interpretable sufficient statistics. We illustrate the OPP by studying US monetary policy decisions.

C0367: **Heterogeneity and aggregate fluctuations**
Presenter: **Minsu Chang**, Georgetown University, United States
Co-authors: Frank Schorfheide, Xiaohong Chen
A state-space model is developed with a state-transition equation that takes the form of a functional vector autoregression and stacks macroeconomic aggregates and a cross-sectional density. The measurement equation captures the error in estimating log densities from repeated cross-sectional samples. The log densities and the transition kernels in the law of motion of the states are approximated by sieves, which leads to a finite-dimensional representation in terms of macroeconomic aggregates and sieve coefficients. We use this model to study the joint dynamics of technology shocks, per capita GDP, employment rates, and earnings distribution. We find that the estimated spillovers between aggregate and distributional dynamics are generally small. A positive technology shock tends to decrease inequality, and a shock that raises the inequality of earnings leads to a small but not significant increase in GDP.

C0370: **Learning about the long run**
Presenter: **Leland Farmer**, University of Virginia, United States
Co-authors: Emi Nakamura, Jon Steinsson

Forecasts of professional forecasters are anomalous: they are biased, forecast errors are autocorrelated, and they are predictable. Sticky or noisy information models seem unlikely explanations for these anomalies: professional forecasters pay attention constantly and have precise knowledge of the data in question. We propose that these anomalies arise because professional forecasters do not know the model that generates the data. We show that Bayesian agents learning about hard-to-learn features of the data generating process (low-frequency behavior) generate all the prominent anomalies emphasized in the literature. We show this for two applications: professional forecasts of nominal interest rates for the sample period 1980-2019 and CBO forecasts of GDP growth for the sample period 1976-2019. Our learning model for interest rates also provides an explanation for deviations from the expectations hypothesis of the term structure that does not rely on time-variation in risk premia.

C1053: **Forecasting with a panel tobit model**
Presenter: **Laura Liu**, Indiana University Bloomington, United States
Co-authors: Roger Moon, Frank Schorfheide

A dynamic panel Tobit model with heteroskedasticity is used to generate point, set, and density forecasts for a large cross-section of short time series of censored observations. Our fully Bayesian approach allows us to flexibly estimate the cross-sectional distribution of heterogeneous coefficients and then implicitly use this distribution as prior to constructing Bayes forecasts for the individual time series. We construct set forecasts that explicitly target the average coverage probability for the cross-section. We present a novel application in which we forecast bank-level charge-off rates for credit card and residential real estate loans, comparing various versions of the panel Tobit model.

C0354: **Local projections vs. VARs: Lessons from thousands of DGPs**
Presenter: **Mikkel Plagborg-Moller**, Princeton University, United States
Co-authors: Christian Wolf, Dake Li

A simulation study is conducted for Local Projection (LP) and Vector Autoregression (VAR) estimators of structural impulse responses across thousands of data generating processes (DGPs), designed to mimic the properties of the universe of U.S. macroeconomic data. The analysis considers various structural identification schemes and several variants of LP and VAR estimators, and we pay particular attention to the role of the researcher's loss function. A clear bias-variance trade-off emerges: Because our DGPs are not exactly finite-order VAR models, LPs have lower bias than VAR estimators; however, the variance of LPs is substantially higher than that of VARs at intermediate or long horizons. Unless researchers are overwhelmingly concerned with bias, shrinkage via Bayesian VARs or penalized LPs is attractive.

CO170 Room Virtual R34 ADVANCES IN FINANCIAL MODELLING AND INFERENCE Chair: Richard Luger

C1143: **A supply and demand approach to equity pricing**
Presenter: **Evan Jo**, Queen's University, Canada
Co-authors: Sebastien Betermier, Laurent Calvet

A tractable general equilibrium framework is developed which provides a direct mapping between (i) the supply and demand for capital at the firm level and (ii) the cross-section of stock returns. Investor behavioural tilts and hedging needs drive capital supply, while firm profitability drives demand. Heterogeneity in supply and demand factors determines the sign of the risk-return relation and generates anomalies such as betting-against-beta, betting-against-correlation, size, value, investment, and profitability. We estimate the supply and demand schedules of over 4,000 U.S. firms and verify that the model accurately predicts the sign of the risk-return relation conditional on characteristics.

C1092: **Unfolded skewness and kurtosis timings in out-of-sample density forecasts of financial returns**
Presenter: **Xiaochun Liu**, University of Alabama, United States

The aim is to evaluate both statistical significance and economic relevance of distributional timings in out-of-sample density forecasts. We estimate a wide range of specifications with dynamic higher moments and conduct a variety of statistical tests. The test results consistently show that modeling time-varying skewness and kurtosis significantly improves both adequacy and accuracy of density forecasts. In the utility-based comparisons, we find that switching to time-varying skewness- and kurtosis-based portfolios from constant-higher-moments-based portfolios yields a gain of an extra 286 and 395 basis points on average per year, respectively. Moreover, an investor is willing to pay 90 and 140 basis points per year to acquire skewness and kurtosis information beyond volatility information. Among the competing models, the unfolded GARCH model, which decomposes returns into the product of their absolute values and signs, presents not only the strongest statistical significance for the left tail of financial returns, but yields the highest gains of about 8.7% and 11.9% on average for an investor unfolding skewness and kurtosis timings.

C1086: **Covariates hiding in the tails**
Presenter: **Lerby Ergun**, Bank of Canada, Canada

Scaling behavior measured in cross-sectional studies through the tail index of a power law is prone to a bias. This hampers inference; in particular, time variation in estimated tail indices may be erroneous. In the case of a linear factor model, the factor biases the tail indices in the left and right tails in opposite directions. This fact can be exploited to reduce bias. We show how this bias arises from the factor, how to remedy for the bias and how to apply our methods to financial data and geographic location data.

C1219: **Evaluating factor strength with a large number of assets**
Presenter: **Sermin Gungor**, Bank of Canada, Canada
Co-authors: Richard Luger

An asymptotic test is developed for evaluating "factor strength" in linear factor pricing models using a large number of assets. Factor models explain the expected return differentials across assets as a linear function of their exposures (betas) to a small number of risk factors. In doing so, they rely on the assumption that factors under consideration are well identified (so-called strong factors). A failure of this identification condition implies that a factor is either "weak" with small betas or "irrelevant" with exactly zero betas. In either case, the estimation of the factor risk premium becomes unreliable with severe overrejection of a zero premium on the weak/irrelevant factor. The proposed methodology identifies weak and irrelevant factors while adjusting the probability of false rejections resulting from a large number of assets. i.e. multiple testing problems. An application using individual stock data for the U.S. stock market finds that the market factor is the only strong factor, but its strength varies over time with the market composition.

C1318: **Regularizing stock return covariance matrices via multiple testing**
Presenter: **Richard Luger**, Laval University, Canada

A new method is developed for the regularization of stock return covariance matrices in well-diversified portfolio settings. The framework is quite general and allows for the presence of heavy tails and multivariate GARCH-type effects of unknown form. The approach proceeds by simultaneously testing all pairwise correlations and then sets to zero the elements that are not statistically significant. Distribution-free Monte Carlo test procedures are proposed for control of the familywise error rate. A subsequent shrinkage step ensures that the covariance matrix estimate is positive definite and well-conditioned, while preserving the achieved zeros. When compared to alternative estimators, the new regularization method is found to perform well both in simulation experiments and in an actual portfolio optimization application with low-volatility stocks.

CO758 Room Virtual R36 TEXT MINING IN ECONOMICS Chair: Paul Hofmarcher

C0720: **Text-based ideal points**
Presenter: **Keyon Vafa**, Columbia University, United States
Ideal point models analyze lawmakers votes to quantify their political positions, or ideal points. But votes are not the only way to express a political position. Lawmakers also give speeches, release press statements, and post tweets. We will present the text-based ideal point model (TBIP), an unsupervised probabilistic topic model that analyzes texts to quantify the political positions of its authors. We demonstrate the TBIP with two types of politicized text data: U.S. Senate speeches and senator tweets. Though the model does not analyze their votes or political affiliations, the TBIP separates lawmakers by party, learns interpretable politicized topics, and infers ideal points close to the classical vote-based ideal points. One benefit of analyzing texts, as opposed to votes, is that the TBIP can estimate ideal points of anyone who authors political texts, including non-voting actors. To this end, we use it to study tweets from the 2020 Democratic presidential candidates. Using only the texts of their tweets, it identifies them along an interpretable progressive-to-moderate spectrum.

C0279: **Adversarial learning of Poisson factorisation models for gauging brand sentiment in user reviews**
Presenter: **Runcong Zhao**, University of Warwick, United Kingdom
The Brand-Topic Model, a probabilistic model which is able to generate polarity-bearing topics of commercial brands, is presented. Compared to other topic models, BMT infers real-valued brand-associated sentiment scores and extracts fine-grained sentiment-topics which vary smoothly in a continuous range of polarity scores. It builds on the Poisson factorisation model, combining it with an adversarial learning mechanism to induce better-separated polarity-bearing topics. Experimental evaluation on Amazon reviews against several baselines shows an overall improvement of topic quality in terms of coherence, uniqueness and separation of polarised topics.

C0806: **Time-evolving text-based ideal point model to infer partisanship in the US senate**
Presenter: **Sourav Adhikari**, Vienna University of Economics and Business, Austria
Co-authors: Bettina Gruen, Paul Hofmarcher
Ideal point models analyze lawmakers' votes, speeches, press statements and social media posts to quantify their political positions along a latent continuum. We extend the text-based ideal point model to obtain a time-evolving version to study the evolution of the ideological positions of lawmakers over time and assess the change in the average difference in bipartisanship among representatives from two political parties. We aim to confirm recent findings regarding the increase in partisanship manifested in speeches by Republicans and Democrats in the US Senate during the last years. These findings were drawn using a penalized estimator for measuring group differences in choices with high dimensional data and text analysis based on manually pre-defined topics. By contrast, the time-evolving text-based ideal point model infers topics in a data-driven way and does not use the known party membership to infer the ideological positions and thus is not susceptible to overrating spurious differences in vocabulary use of different party members. Drawing the same substantive conclusions based on the results of two different statistical text analysis methods provides evidence for their robustness.

C0247: **Real oil price forecasting: Gains and pitfalls of text data**
Presenter: **Luigi Gifuni**, University of Glasgow, United Kingdom
New text-based measures are developed for assessing human sentiment and economic uncertainty in the oil market. Empirical experiments show that sentiment indexes are very responsive to historical geopolitical events that have affected the price of oil. In contrast, uncertainty indicators may hide structural pitfalls, which create problems when alternative measures of the real oil price are forecasted. We propose a linear kernelization of output forecasts, in order to achieve the best forecasting performance at any time horizon.

C0827: **Assessing the effects of friend-to-friend texting on turnout in the 2020 U.S. presidential election**
Presenter: **Aaron Schein**, Columbia University, United States
Political campaigns in recent elections have started to embrace friend-to-friend organizing, in which volunteers organize and encourage their own close contacts to cast a ballot on Election Day. The premise of friend-to-friend organizing is that get out the vote (GOTV) encouragements are more effective when delivered by trusted messengers, like friends or family members, than when delivered by strangers, as is the case with traditional GOTV tactics like door-to-door canvassing or phone-banking. This is among the first large-scale experimental studies to assess the causal effects of friend-to-friend GOTV tactics. We find small treatment effects of friend-to-friend text-message reminders using data from the 2020 US election.

CO408 Room Virtual R38 SENTOMETRICS Chair: David Ardia

C0457: **Weekly economic index for Belgium: Design and validation**
Presenter: **Arno De Block**, Vrije Universiteit Brussel, Belgium
Co-authors: Feliciaan De Palmenaer, Kris Boudt
GDP growth is the key indicator for measuring the present state of the economy. Policymakers, firms, and investors closely monitor it to optimize their economic decision-making. But while decisions must be made in real-time, official GDP figures are only observed quarterly and with a substantial publication lag. Exploiting timely available data, one can predict the present state of the economy using econometric models, often referred to as nowcasting. The goal is to create a nowcast for Belgium on a weekly basis that tracks the economic activity. This Belgian Weekly Economic Index combines three data sources: banking data, electricity load data and media sentiment data. The banking data consists of sectoral aggregates of in- and outgoing cashflows of small to medium companies. The industry mainly dominates the electricity load, while economic sentiment can be extracted from newspaper articles. Combining these heterogeneous data sources using econometric models will lead to a weekly tracker of the Belgian economy.

C0465: **Aggregation of bank transaction data into a weekly Index of economic activity**
Presenter: **Feliciaan De Palmenaer**, Universiteit Gent and Vrije Universiteit Brussel, Belgium
Co-authors: Milan van den Heuvel, Kris Boudt, Koen Schoors
Bank transaction data gives a high-frequency overview of monetary transfers between firms, aggregating certain transactions, the output of a firm can be constructed for a week. When aggregating all firms over a week, an index can be constructed to show the week-over-week growth or changes in economic activity of the Belgian economy. The anonymized dataset contains all bank transactions of firms that are clients of a Belgian bank. Going from raw transactions data to the weekly index is a multi-step process with some challenges. Aggregating is done in three steps, first as much as possible, filtering, seasonal adjustments and cleaning is made on a per firm basis. The next step is to group all firms in the same sector, using NACE-codes, an EU wide classification nomenclature, and take the mean of all firms' outputs. The final step is to aggregate these sectoral indices to one weekly economic index. The first correction we had to on our data was to add sectoral seasonality adjustments to firms' outputs to prevent cyclical deviations, using the geometric mean of the weekly outputs of firms in this sector for 10 years. Instead of weeks, pseudo-weeks are used where each month consists out of 4 weeks. To correct for the longer weeks, all weekly data is normalized. And to prevent differences between the weight of a sector in the overall economy and the weight in the bank index, correct weights were used from the Belgian statistical office when constructing the index.

C0646: **Forecasting GDP in Europe with textual data**
Presenter: **Luca Barbaglia**, European Commission Joint Research Centre, Italy
Co-authors: Sergio Consoli, Sebastiano Manzan
The aim is to evaluate the informational content of news-based sentiment indicators for forecasting the Gross Domestic Product (GDP) of the five major European economies. The sentiment indicators that we construct are aspect-based, in the sense that we consider only the text that is related to a specific economic aspect of interest. In addition, the sentiment is fine-grained as each word is assigned a score in the interval [-1, 1]. Our data set includes over 27 million articles for 26 major newspapers in 5 different languages. The evidence indicates that these sentiment indicators are significant predictors to forecast GDP and their predictive content is robust to controlling for macroeconomic and survey confidence indicators available to forecasters in real-time. We also discuss the application of the sentiment indicators during the Covid-19 pandemic and demonstrate their relevance in nowcasting GDP.

C1405: **Sentopics: An R package linking topic models and sentiment analysis**
Presenter: **Olivier Delmarcelle**, Ghent University, Belgium
The R package sentopics provides a new approach to analyse textual sentiment. By integrating topic modelling in the framework, the package breaks down classical sentiment values into topical-sentiment components. The package implements the Joint Sentiment-Topic model, a two-layer topic model incorporating sentiment lexicons to determine topical-sentiment word clusters. Sentopics also enables to join the results of a simple topic model with sentiment computed from another tool. In addition, the package defines several functions designed to enhance interpretability by vizualizing topic models results or preparing topical-sentiment time series. Finally, a built-in parallel framework is included to speed up the estimation of multiple topic models at once. The user may then compare the different estimates using coherence metrics, a tool implemented by the package to assess the quality of topic models.

C1782: **Climate change concerns and the performance of green versus brown stocks**
Presenter: **David Ardia**, HEC Montreal, Canada
The aim is to empirically test the previous prediction that green firms outperform brown firms when concerns about climate change increase unexpectedly, using data for S&P 500 companies from January 2010 to June 2018. To capture unexpected increases in climate change concerns, we construct a Media Climate Change Concerns index using news about climate change published by major U.S. newspapers. We find that when concerns about climate change increase unexpectedly, green firms stock prices increase, while brown firms decrease. Further, using topic modeling, we conclude that climate change concerns affect returns both through investors updating their expectations about firms future cash flows and through changes in investors preferences for sustainability.

CO698 Room K2.31 Nash (Hybrid 07) FACTOR MODELS IN ASSET PRICING Chair: Svetlana Bryzgalova

C1456: **Asset pricing with missing data**
Presenter: **Markus Pelger**, Stanford University, United States
Co-authors: Svetlana Bryzgalova, Martin Lettau, Sven Lerner
The aim is to show how to impute missing information for stock returns and to study the implications for asset pricing relative to the current standard of using only observed or ad-hoc imputed values. Missing data in firm characteristics is a prevalent problem. As firm characteristics are not missing at random, using only observed data for building or evaluating asset pricing model results in biased estimates. We exploit the dependency of firm-characteristics in their time and in the cross-sectional dimension to impute the missing values for a large dimensional cross-section. Our imputed values are a substantial improvement relative to ah-hoc procedures as for example simple cross-sectional averages or past observations. In a large scale empirical analysis we study the asset pricing implications.

C1459: **Interpretable deep learning in asset pricing**
Presenter: **Andreas Neuhierl**, Washington University in St. Louis, United States
Co-authors: Jianqing Fan, Tracy Ke, Yuan Liao
A new nonparametric methodology is developed for estimating conditional asset pricing models using deep neural networks. We employ time-varying conditional information on betas carried by firm-specific characteristics. The method first applies cross-sectional deep learning, period-by-period to estimate spontaneous conditional expected returns, defined as the conditional expectation of asset returns given characteristics and factor realizations. We also estimate the long-term expected return as the predicted mispricing component and the product of the estimated risk exposures times the price of risk. We apply local kernel smoothing to capture the return dynamics that arise from time-varying alphas and betas. We formally establish the asymptotic theory of the deep-learning estimators for conditional expected returns, alphas and risk premia, which apply to both in-sample fit and out-of-sample predictions. Contrary to many applications of neural networks in economics, we can open the black box and provide an economic interpretation of the successful predictions obtained from neural networks. We decompose predicted returns into a risk-based and mispricing component. Empirically, we find a large, time-varying mispricing component. We find that the mispricing component is slowly decaying over time, but not monotonically. Mispricing tends to be high during times of high market volatility which is linked to periods of economic turmoil.

C1475: **Factor models for conditional asset pricing**
Presenter: **Paolo Zaffaroni**, Imperial College London, United Kingdom
A methodology is developed for inference on conditional asset pricing models robust to omitted risk factors and to misspecified conditional dynamics. All the features of the asset pricing model, such as risk premia, factors exposures, factors variances and covariances, idiosyncratic risk, and number of risk factors, are potentially time-varying. The limiting results hold when the number of assets diverges but the time-series dimension is fixed, possibly very small, applicable to a variety of data frequencies. An extensive empirical application based on individual asset returns data demonstrates the powerfulness of the methodology, allowing to tease out the empirical content of the time-variation elicited by asset pricing theory.

C1550: **On the power of asset pricing tests**
Presenter: **Jiacui Li**, University of Utah, United States
Co-authors: Cesare Robotti
Researchers often disagree on how to interpret the evidence from asset pricing tests. For instance, Fama and French three-factor model was shown to help to explain size and book-to-market related return predictability. However, it was later argued that such evidence has low power against the characteristic-based (anomaly) view. To contribute to resolving these debates, we propose a simulation-based approach to benchmark the evidence from asset pricing tests. Specifically, we apply asset pricing tests to simulated characteristics that predict returns as anomalies. The simulations reveal that tests that employ the same characteristics to form factors and test assets often show nontrivial explanatory power for mechanical reasons, and the results obtained in the original papers are based on methods that do not enable us to reject the null hypothesis that all characteristics are anomalies. Tests, where the factors are formed using different characteristics, are often not mechanical. The recently proposed instrumented principal component analysis test (IPCA) has more power to differentiate between factors and anomalies. This simulation-based method is easy to interpret, easy to implement, and can flexibly accommodate different null hypotheses. Overall, our findings imply that proper benchmarking can help us better interpret the evidence from asset pricing tests.

C1579: **Smart stochastic discount factors**
Presenter: **Sofonias Alemu Korsaye**, University of Geneva, Ethiopia
Co-authors: Fabio Trojani, Alberto Quaini

A novel no-arbitrage framework is proposed which exploits convex asset pricing constraints to study investors' marginal utility of wealth or, more generally, Stochastic Discount Factors (SDFs). We establish a duality between minimum dispersion SDFs and penalized portfolio selection problems, building the foundation for characterizing the feasible tradeoffs between a SDF's pricing accuracy and its comovement with systematic risks. Empirically, a minimum variance CAPM-SDF produces a Pareto optimal tradeoff. This SDF only depends on two distinct risk factors: A traded market factor and a minimum variance excess return that bounds the mispricing of risks unspanned by market shocks.

CO328 Room K2.41 (Hybrid 09) PREDICTIVE MODELLING OF FINANCIAL DATA Chair: Robert Taylor

C0202: **Extensions to IVX methods of inference for return-predictability**
Presenter: **Robert Taylor**, University of Essex, United Kingdom
Co-authors: Matei Demetrescu, Paulo Rodrigues, Iliyan Georgiev

IVX methods have proved particularly valuable in predictive regressions as they allow for possibly strongly persistent and endogenous regressors. We make three contributions. First, we demonstrate that, provided either a suitable bootstrap implementation is employed or heteroscedasticity-consistent standard errors are used, previous IVX-based predictability tests retain asymptotically pivotal inference, regardless of the degree of persistence or endogeneity of the (putative) predictor, under considerably weaker assumptions on the innovations than were required formerly. Second, we develop asymptotically valid bootstrap implementations of the IVX tests under these conditions. Monte Carlo simulations show that the bootstrap methods we propose can deliver considerably more accurate finite-sample inference than the asymptotic implementation of these tests used previously under certain problematic parameter constellations, most notably for their implementation against one-sided alternatives, and where multiplepredictors are included. Third, under the same conditions as we consider for the full sample tests, we show how sub-sample implementations of the IVX approach, coupled with a suitable bootstrap, can be used to develop asymptotically valid one-sided and two-sided tests for the presence of temporary windows of predictability.

C0206: **Testing the predictability of stock returns with smoothly varying deterministic mean**
Presenter: **Matei Demetrescu**, CAU Kiel, Germany
Co-authors: Mehdi Hosseinkouchak

Checking whether stock returns may be predicted using financial valuation ratios and other fundamental values is facing several methodological and empirical challenges. Most importantly, the typical putative predictor variable exhibits high persistence, which leads to nonstandard limiting distributions of the OLS estimator and associated t statistic in predictive regressions. While there are several methods to deal with the issue of nonstandard distributions, the high predictor persistence also opens the door to spurious regression findings induced by time-varying mean components of stock returns, if not properly controlled for. Such control requires additional information, which may not be available in practice. We take a different approach and robustify IVX predictive regression to the presence of smooth trend components. To this end, we employ a particular local mean adjustment scheme to account for possibly time-varying means. The limiting distribution of the resulting IVX t statistic is derived under sequences of local alternatives, and a wild bootstrap implementation improving the finite-sample behavior is provided. Compared to IVX predictive regression, there is a price to pay for robustness in terms of power; at the same time, the IVX statistic without adjustment consistently rejects the false null of no predictability in the presence of ignored time-varying deterministic mean components.

C0604: **Out of sample predictability in predictive regressions with many predictor candidates**
Presenter: **Jean-Yves Pitarakis**, University of Southampton, United Kingdom
Co-authors: Jesus Gonzalo

The focus is on detecting the presence of out of sample predictability in linear predictive regressions with a potentially large set of candidate predictors. We propose a procedure based on out of sample MSE comparisons that is implemented in a pairwise manner using one predictor at a time and resulting in an aggregate test statistic that is standard normally distributed under the null hypothesis of no linear predictability. Predictors can be highly persistent, purely stationary or a combination of both. Upon rejection of the null hypothesis, we subsequently introduce a predictor screening procedure designed to identify the most active predictors.

C1129: **Tests for time series momentum**
Presenter: **Paulo Rodrigues**, Universidade Noval de Lisboa, Portugal
Co-authors: Matei Demetrescu, Robert Taylor

Cross-sectional momentum is a well-known asset-pricing anomaly. Consistent evidence of a new anomaly across different asset classes and markets has been found which has been termed time-series momentum. Specifically, they find that the past 12-month excess returns of an asset are a good predictor of its future. However, this conclusion is not yet consensual. For instance, standard univariate and pooled panel predictive regressions has been recently used to find little evidence of time-series momentum in the same set of assets analysed previously. These contrasting results highlight the importance of rigorous inference on the time series predictability of future returns based on past cumulative returns when testing for this phenomenon. We evaluate the performance of existing methods and propose new solutions for testing for this anomaly.

C1505: **The role of investor sentiment in commodity price behaviour: some evidence via time-varying causality tests**
Presenter: **Roderick McCrorie**, University of St Andrews, United Kingdom
Co-authors: Mario Lupoli

A standard Granger-causality test has been previously used to provide evidence of causality from index investment to non-ferrous metals and some agricultural commodity prices. The results support that money flows associated with index investors helped predict changes in crude oil futures prices. We examine the robustness of such results using a more powerful and encompassing causality testing strategy based on another test which allows the causality measure to be time-varying. This helps gauge the magnitude of the causal relation from index investment to log-returns and pinpoints when it is statistically significant. We confirm the key finding of causality, albeit in a milder form, from index investment to some non-ferrous metals and agricultural commodities but find the causal relation is measured significantly only around the Global Financial Crisis and other identifiable dates. This highlights that standard causality tests can be affected by extreme price behaviour and that apposite testing strategies should be used. A granularity is offered that helps clarify an academic literature that has been settling against the efficacy and wider applicability of the Singleton result but has not proved persuasive enough to see the same view become pervasive among finance practitioners.

Saturday 18.12.2021 18:45 - 20:00 Parallel Session F – CFE-CMStatistics

EO714 Room K0.16 (Hybrid 02) BAYESIAN MODEL SELECTION Chair: Jairo Fuquene

E0394: **Spike-and-slab group lassos for grouped regression and sparse generalized additive models**
Presenter: **Ray Bai**, University of South Carolina, United States

The spike-and-slab group lasso (SSGL) is introduced for Bayesian estimation, and variable selection in linear regression with grouped variables is introduced. We further extend the SSGL to sparse generalized additive models (GAMs), thereby introducing the first nonparametric variant of the spike-and-slab lasso methodology. The model simultaneously performs group selection and estimation. At the same time, our fully Bayes treatment of the mixture proportion allows for model complexity control and automatic self-adaptivity to different levels of sparsity. We develop theory to uniquely characterize the global posterior mode under the SSGL and introduce a highly efficient block coordinate ascent algorithm for maximum a posteriori (MAP) estimation. We further employ de-biasing methods to provide uncertainty quantification of our estimates. Thus, implementation of our model avoids the computational intensiveness of Markov chain Monte Carlo (MCMC) in high dimensions. We derive posterior concentration rates for both grouped linear regression and sparse GAMs when the number of covariates grows at nearly exponential rate with sample size. Finally, we illustrate our methodology through extensive simulations and data analysis.

E0631: **Dynamics of mean-field approximation: A case-study in singular models**
Presenter: **Anirban Bhattacharya**, Texas AM University, United States
Co-authors: Debdeep Pati, Yun Yang, Sean Plummer

The marginal likelihood or evidence in Bayesian statistics contains an intrinsic penalty for larger model sizes and is a fundamental quantity in Bayesian model comparison. Over the past two decades, there has been steadily increasing activity to understand the nature of this penalty in singular statistical models, building on pioneering work by Sumio Watanabe. Unlike regular models where the Bayesian information criterion (BIC) encapsulates a first-order expansion of the logarithm of the marginal likelihood, parameter counting gets trickier in singular models where a quantity called the real log canonical threshold (RLCT) summarizes the effective model dimensionality. We show that mean-field variational inference correctly recovers the RLCT for any singular model in its canonical or normal form. We additionally exhibit the sharpness of our bound by analyzing the dynamics of a general-purpose coordinate ascent algorithm (CAVI) popularly employed in variational inference.

E1069: **Using log Cauchy priors for modeling sparsity**
Presenter: **Xueying Tang**, University of Arizona, United States
Co-authors: Zihan Zhu

Sparsity is often a desired structure for parameters in high-dimensional statistical problems. Within a Bayesian framework, sparsity is usually induced by spike-and-slab priors or global-local shrinkage priors. The latter choice is often expressed as a scale mixture of normal distributions. It marginally places a polynomial-tailed distribution on the parameter. In general, a heavier-tailed distribution has a better performance in estimating sparse parameters. We consider the log Cauchy priors in the normal mean estimation problem. This class of priors is proper while having a tail order arbitrarily close to one. The resulting posterior mean is a shrinkage estimator, and the posterior contraction rate is sharp minimax. We will also demonstrate these theoretical properties through simulations.

EO456 Room K0.19 (Hybrid 04) ADVANCES IN CAUSAL INFERENCE (VIRTUAL) Chair: Luke Keele

E1146: **Global and local estimators of a dose-response curve**
Presenter: **Matteo Bonvini**, Carnegie Mellon University, United States
Co-authors: Edward Kennedy

The problem of estimating a dose-response curve, both globally and locally at a point, is considered. Letting A denote a continuous treatment variable, the target of inference is the expected outcome if everyone in the population takes treatment level $A = a$. Under standard assumptions, the dose-response function takes the form of a partial mean. Building upon the recent literature on nonparametric regression with orthogonal signals, we study three different estimators. As a global method, we build upon previous work to construct an empirical-risk-minimization-based estimator and give an explicit characterization of second-order remainder terms. As a local method, we develop a DR-learner. Finally, we construct an mth order estimator based on the theory of higher-order influence functions. For each estimator, we provide an upper bound on the mean-square error and investigate its finite-sample performance in a simulation.

E1140: **A negative correlation strategy for bracketing in difference-in-differences**
Presenter: **Luke Keele**, University of Pennsylvania, United States
Co-authors: Ting Ye

The method of difference-in-differences (DID) is widely used to study the causal effect of policy interventions in observational studies. DID employs a before and after comparison of the treated and control units to remove bias due to time-invariant unmeasured confounders under the parallel trends assumption. Estimates from DID, however, will be biased if the outcomes for the treated and control units evolve differently in the absence of treatment, namely if the parallel trends assumption is violated. We propose a general identification strategy that leverages two groups of control units whose outcomes relative to the treated units exhibit a negative correlation, and achieves partial identification of the average treatment effect for the treated. The identified set is of a union bounds form that involves the minimum and maximum operators, which makes the canonical bootstrap generally inconsistent and naive methods overly conservative. By utilizing the directional inconsistency of the bootstrap distribution, we develop a novel bootstrap method to construct uniformly valid confidence intervals for the identified set and parameter of interest when the identified set is of a union bounds form, and we establish the method's theoretical properties. We develop a simple falsification test and sensitivity analysis. We apply the proposed strategy for bracketing to study whether minimum wage laws affect employment levels.

E1173: **Missing, presumed different: Quantifying the risk of attrition bias in education evaluations**
Presenter: **Ben Weidmann**, Harvard Kennedy School, United Kingdom

The magnitude of attrition bias for 10 randomized controlled trials (RCTs) in education is estimated. We make use of a unique feature of administrative school data in England that allows us to analyse post-test academic outcomes for nearly all students, including those who originally dropped out of the RCTs we analyse. We find that the typical magnitude of attrition bias is 0.015 effect size units (ES), with no estimate greater than 0.034 ES. This suggests that, in practice, the risk of attrition bias is limited. However, this risk should not be ignored as we find some evidence against the common Missing At Random assumption. Attrition appears to be more problematic for treated units. We recommend that researchers incorporate uncertainty due to attrition bias, as well as perform sensitivity analyses based on the types of attrition mechanisms that are observed in practice.

EO320 Room Virtual R18 INTERPRETABILITY AND TRUSTWORTHINESS IN MACHINE LEARNING Chair: Efstathios Gennatas

E1327: **Conservative approaches to integrative high-dimensional studies of the brain and brain-related measurements**
Presenter: **Brian Avants**, University of Virginia and Invicro, United States

Wide data collection of neuroimaging and brain-related measurements is increasingly easy in both research and clinical setting. However, this leads to several issues: siloing of analyses, p-hacking and other statistical sins that may inadvertently promote irreproducible results. We describe how we use pre-defined integrative statistical models to help reduce multiple modalities to much smaller spaces, identify the reliable signal in

datasets with relatively few subjects relative to the number of predictors and how we interpret these results using domain knowledge to enhance systems-level understanding of the brain.

E1749: **Stability-driven interpretation of deep learning models: A neuroscience case study**
Presenter: **Reza Abbasi-Asl**, University of California, San Francisco, United States

In the past decade, research in machine learning has been exceedingly focused on the development of algorithms and models with remarkably high predictive capabilities. These predictive models have wide applications in large-scale data-driven domains including neuroscience, healthcare, and computer vision. However, interpreting these models still remains a challenge, primarily because of the large number of parameters involved. We will introduce two frameworks based on (1) stability and (2) compression to build more interpretable machine learning models. These two frameworks will be demonstrated in the context of a computational neuroscience study. First, we will introduce a stability-driven visualization framework for models based on neural networks. This framework is successful in characterizing complex biological neurons in the mouse and non-human primate visual cortex. This visualization uncovers the diversity of stable patterns explained by neurons. Then, we will discuss two neural network compression techniques based on iterative pruning and low dimensional decomposition of filters. These model compression techniques increase the interpretability of networks while retaining the high accuracy and diversity of filters. The compressed models give rise to a new set of accurate models for neurons but with much simpler structures.

E1750: **Friends do not let friends deploy black-box models: The importance of intelligibility in machine learning in healthcare**
Presenter: **Rich Caruana**, Microsoft Research, United States

In machine learning sometimes a tradeoff must be made between accuracy and intelligibility: the most accurate models often are black-box models that are not very intelligible, and the most intelligible models usually are less accurate. This can limit the accuracy of models that can safely be deployed in mission-critical applications such as healthcare where being able to understand, validate, edit, and ultimately trust a model is important. We have developed a learning method based on generalized additive models (GAMs) that is as accurate as full complexity models such as neural nets, boosted trees and random forests, but more intelligible than linear models. This makes it easy to understand what models have learned and to edit models when they learn inappropriate things. Making it possible for medical experts to understand and repair a model is critical because most clinical data is complex and has unanticipated problems. We will present several healthcare case studies where these high-accuracy GAMs discover surprising patterns in the data that would have made deploying black-box models risky. The case studies include surprising findings in pregnancy, pneumonia, ICU and COVID-19 risk prediction.

EO140 Room Virtual R20 STATISTICAL INFERENCE WITH DEEP LEARNING Chair: Faming Liang

E0559: **Variational inference for sparse deep learning**
Presenter: **Qifan Song**, Purdue University, United States

Sparse deep learning aims to address the challenge of huge storage consumption by deep neural networks, and to recover the sparse structure of target functions. Although tremendous empirical successes have been achieved, most sparse deep learning algorithms are lacking theoretical support. On the other hand, another line of works has proposed theoretical frameworks that are computationally infeasible. We train sparse deep neural networks with a fully Bayesian treatment under spike-and-slab priors, and develop a set of computationally efficient variational inferences via continuous relaxation of Bernoulli distribution. The variational posterior contraction rate is provided, which justifies the consistency of the proposed variational Bayes method. Notably, our empirical results demonstrate that this variational procedure provides uncertainty quantification in terms of Bayesian predictive distribution and is also capable to accomplish consistent variable selection by training a sparse multi-layer neural network.

E0564: **A kernel-expanded stochastic neural network**
Presenter: **Faming Liang**, Purdue University, United States
Co-authors: Yan Sun, Faming Liang

The deep neural network suffers from many fundamental issues in machine learning. For example, it often gets trapped into a local minimum in training, and its prediction uncertainty is hard to be assessed. To address these issues, we propose the so-called kernel-expanded stochastic neural network (K-StoNet) model, which incorporates support vector regression (SVR) as the first hidden layer and reformulates the neural network as a latent variable model. The former maps the input vector into an infinite-dimensional feature space via a radial basis function (RBF) kernel, ensuring the absence of local minima on its training loss surface. The latter breaks the high-dimensional nonconvex neural network training problem into a series of low-dimensional convex optimization problems, and enables its prediction uncertainty easily assessed. The K-StoNet can be easily trained using the imputation-regularized optimization (IRO) algorithm. Compared to traditional deep neural networks, K-StoNet possesses a theoretical guarantee to asymptotically converge to the global optimum and enables the prediction uncertainty easily assessed. The performances of the new model in training, prediction and uncertainty quantification are illustrated by simulated and real data examples.

E1176: **An adaptively weighted stochastic gradient MCMC algorithm for monte carlo simulation and global optimization**
Presenter: **Wei Deng**, Purdue University, United States
Co-authors: Guang Lin, Faming Liang

An adaptively weighted stochastic gradient Langevin dynamics (AWSGLD) algorithm is proposed for Bayesian learning of big data problems. The proposed algorithm is scalable and possesses a self-adjusting mechanism: It adaptively flattens the high energy region and protrudes the low energy region during simulations such that both Monte Carlo simulation and global optimization tasks can be greatly facilitated in a single run. The self-adjusting mechanism enables the proposed algorithm to be essentially immune to local traps. Theoretically, by showing the stability of the mean-field system and verifying the existence and regularity properties of the solution of the Poisson equation, we establish the convergence of the AWSGLD algorithm, including both the convergence of the self-adapting parameters and the convergence of the weighted averaging estimators. Empirically, the AWSGLD algorithm is tested on multiple benchmark datasets including CIFAR100 and SVHN for both optimization and uncertainty estimation tasks. The numerical results indicate its great potential in Monte Carlo simulation and global optimization for modern machine learning tasks.

EO242 Room Virtual R21 ADVANCES IN STATISTICAL METHODS FOR MOBILE HEALTH Chair: Walter Dempsey

E1331: **Assessing timing of microrandomized trials using intensive longitudinal data and differential time-varying effect models**
Presenter: **Nicholas Jacobson**, Dartmouth College, United States
Co-authors: Kejin Wu

In the behavioral sciences, methods for delivering interventions within the context of daily life are developing rapidly, fueled by the development of microrandomized controlled trials and experience sampling. Although intensive longitudinal data are often collected to evaluate the immediate effects of these interventions, the timing of these interventions on behavior has been given limited attention. Given this, the field could benefit from a tool to detect and estimate the time in which interventions have an impact on their respective outcomes. Nevertheless, existing tools have difficulty in estimating the timing of interventions. Consequently, we propose an extension of the Differential Time-Varying Effect Model (DTVEM) which attempts to detect the timing of interventions on outcomes by trying to detect the lag intervals between exogenous variables (i.e. intervention delivery) and outcomes. We extend the DTVEM by pairing generalized additive mixed models with linear mixed models to identify optimal time lags and intervention effects. By intensive simulations based on, the efficiency of the DTVEM with additional stage is tested, and the results showed

promising power and point estimates, and low type I error. Consequently, the extended DTVEM allows researchers to perform power analyses regarding the timing of intervention effects and detect the timing of intervention effects using intensive longitudinal data and microrandomized controlled trials.

E1586: **Assessing how well RL algorithms personalize in mobile health**
Presenter: **Xiang Meng**, Harvard University, United States
Co-authors: Zeyang Jia, Peng Liao, Kelly Zhang, Susan Murphy

Reinforcement learning (RL) algorithms are increasingly used in mobile health applications. By adaptively personalizing the delivery of interventions to users, RL algorithms are designed to learn which intervention strategies are most effective in different circumstances for improving users' health outcomes. Therefore, after using an RL algorithm in a mobile health study, it is critical to assess how well the RL algorithm is at personalizing interventions to users: Does it learn over time and achieve better outcomes for users than a non-RL algorithm does? To answer this question, we formally define the meaning of personalization in mobile health and propose statistical methods for investigating whether an algorithm personalizes. We also apply our method to data from a mobile health clinical trial and assess the personalization of the algorithm in the study.

E1587: **Using offline policy evaluation to advance methodological barriers in digital health**
Presenter: **Jane Kim**, Stanford University, United States

Evaluating the efficacy of mHealth apps which deploy algorithms remains a critical priority for health care providers and patients who need this evidence in advance of offering and receiving such care. Contextual bandits are well-suited to address concerns regarding the adaptability of interventions and non-adherence in behavioral interventions; prior work has demonstrated the benefits of deploying bandits in real settings. A key consideration, however, is that it is not usually feasible nor ethical to deploy a new algorithm live in the context of testing health interventions. We will present offline policy evaluation (OPE) as a means to answer practical questions that investigators and intervention designers may have about design questions related to research. OPE can be used to evaluate the performance of algorithms using logged data that was generated under a different policy. Its core idea is to use importance sampling that allows the estimation of a function of a distribution when the data of interest were generated from an altogether different distribution. First, we introduce the method of offline policy evaluation and provide background information about sequential decision making. Then we illustrate its utility using a case example of a digital recommendation system that delivers messages to reduce stress.

EO214 Room Virtual R22 BELIEFS, RISK AND UNCERTAINTY IN ARTIFICIAL INTELLIGENCE I Chair: Davide Petturiti

E0769: **Can probability coherence correction be helpful in data lakes?**
Presenter: **Andrea Capotorti**, Universita degli Studi di Perugia, Italy
Co-authors: Marco Baioletti

Integration, merging, and fusion of uncertain data are more and more current in artificial intelligence since the Big Data era we are leaving in. With the increase of volume of information, the grouping, simplification, and partiality of the statistical analysis become a mandatory task. And with diffuse and duplicated data structures, like e.g. data lakes used for classifications and/or previsions, the problem of inconsistent conditional probabilities assessments comes to the fore. Probabilistic reasoning embedded in a coherent setting has the advantage of formalizing partial knowledge expressed through events, conditional or unconditional, representing different combinations of elementary situations, that can be hence thought of as macro" situations with non-trivial interconnections, implications, incompatibilities, etc., hence it is apt to merge different sources of information. Recently a procedure, based on L1 distance minimization and mixed-integer programming MIP, has been proposed to: correct straight unconditional assessments; revise the belief; solve the statistical matching problem; minimize the number modifications in line with the principle of optimal corrective explanation. Now, an unsupervised revision of incoherent conditional assessments is proposed by profiting from the so-called exploiting of zero probabilities and through shrewd use of slake variables.

E0820: **Incorporating partial prior information in an inferential model**
Presenter: **Leonardo Cella**, North Carolina State University, United States

Even when prior information is available in practice, it is often partial/incomplete, i.e., it does not come in the form of a fully specified probability distribution. Of course, incomplete prior information can mean many different things. On one end of the spectrum, it could be a hard/definitive constraint on some parameter of interest; on the other end, it could be structural assumptions like those commonly assumed in high-dimensional settings (e.g., sparsity); and in the middle, it could be something like a subject-matter expert saying that she is "90% sure that the parameter is in the interval [7,10]" We argue that such priors can be represented without embellishment by random sets, and explore their incorporation within the Inferential Models framework, an originally prior-free approach for valid statistical inference. This incorporation is guided by desired properties, such as that validity is maintained regardless of the truthfulness of the partial prior and that correct partial priors should result in more efficient inferences.

E1043: **Coherence and correction of belief and necessity assessments**
Presenter: **Davide Petturiti**, University of Perugia, Italy
Co-authors: Barbara Vantaggi

Belief functions are uncertainty measures simultaneously close to finitely additive probability measures in terms of properties and able to deal with ambiguity. The elicitation of a belief assessment on a set of events without a particular algebraic structure requires a suitable notion of coherence. De Finetti's work on finitely additive coherent probabilities has shown three possible approaches to coherence: coherence as a consistency notion, coherence as a fair betting scheme, and coherence in terms of penalty criterion. We present generalized notions of coherence for a belief assessment in all the forms recalled above, and prove their equivalence. In turn, this allows us to single out different interpretations of a belief assessment. We further introduce the specialization of such conditions to work in the sub-framework of finitely minitive necessity measures. Finally, we propose a general penalty criterion based on a (strictly) proper scoring rule and investigate the correction of an incoherent assessment in the different frameworks.

EO599 Room Virtual R23 STATISTICAL METHODS FOR ENVIRONMENTAL MIXTURES Chair: Glen McGee

E0461: **Group inverse-gamma gamma shrinkage for sparse regression with application to correlated environmental exposure data**
Presenter: **Jonathan Boss**, University of Michigan, United States
Co-authors: Jyotishka Datta, Xin Wang, Sung Kyun Park, Jian Kang, Bhramar Mukherjee

Heavy-tailed continuous shrinkage priors, such as the horseshoe prior, are widely used for sparse estimation problems. However, there is limited work extending these priors to incorporate bi-level shrinkage for predictors with grouping structures explicitly. Regression coefficient estimation is particularly interesting, where pockets of high collinearity in the covariate space are contained within known covariate groupings. To assuage variance inflation due to multicollinearity, we propose the group inverse-gamma gamma (GIGG) prior, a heavy-tailed prior that can trade-off between local and group shrinkage in a data-adaptive fashion. A special case of the GIGG prior is the group horseshoe prior, whose shrinkage profile is correlated within-group such that the regression coefficients marginally have exact horseshoe regularization. We show posterior consistency for regression coefficients in linear regression models and posterior concentration results for mean parameters in sparse normal means models. The full conditional distributions corresponding to GIGG regression can be derived in closed form, leading to straightforward posterior computation. We show that GIGG regression results in low mean-squared error across a wide range of correlation structures and within-group signal densities

via simulation. We apply GIGG regression to data from the National Health and Nutrition Examination Survey for associating environmental exposures with liver functionality.

E0617: **Bayesian multiple index models for environmental mixtures**
Presenter: **Glen McGee**, University of Waterloo, Canada
Co-authors: Ander Wilson, Thomas Webster, Brent Coull

An important goal of environmental health research is to assess the risk posed by mixtures of environmental exposures. Two popular classes of models for mixtures analyses are response-surface methods and linear-index methods. Response-surface methods estimate high-dimensional surfaces and are highly flexible but difficult to interpret. Linear-index methods decompose coefficients from a linear model into an overall mixture effect and component weights; these models yield easily interpretable effect estimates and efficient inferences but can be overly restrictive. We propose a Bayesian multiple index model framework that combines the strengths of each, allowing for non-linear and non-additive relationships between exposure indices and a health outcome, while reducing dimensionality and estimating index weights. The proposed framework allows one to select an appropriate analysis from a spectrum of models varying in flexibility and interpretability, and it contains both response-surface and linear-index models as special cases. Unlike fully non-parametric alternatives, the framework also provides a means of incorporating prior knowledge about mixtures in future analyses.

E0815: **Estimating perinatal critical windows to environmental mixtures via structured Bayesian regression tree-pairs**
Presenter: **Ander Wilson**, Colorado State University, United States
Co-authors: Daniel Mork

Maternal exposure to environmental chemicals during pregnancy can alter birth and children's health outcomes. The aim is to identify critical windows, time periods when the exposures can change future health outcomes, and estimate the exposure-response relationship. Existing statistical approaches focus on estimation of the association between maternal exposure to a single environmental chemical observed at high-temporal resolution, such as weekly throughout pregnancy, and children's health outcomes. Extending to multiple chemicals observed at high temporal resolution poses a dimensionality problem and statistical methods are lacking. We propose a tree-based model for mixtures of exposures that are observed at high temporal resolution. The proposed approach uses an additive ensemble of structured tree-pairs that define structured main effects and interactions between time-resolved predictors and variable selection to select out of the model predictors not correlated with the outcome. We apply our method in a simulation and the analysis of the relationship between five exposures measured weekly throughout pregnancy and resulting birth weight in a Denver, Colorado birth cohort. We identified critical windows during which fine particulate matter, sulfur dioxide, and temperature are negatively associated with birth weight and interaction between fine particulate matter and temperature. Software is made available in the R package dlmtree.

EO142 Room Virtual R24 RECENT DEVELOPMENTS IN CHANGE-POINT DETECTION METHODS Chair: Shahina Rahman

E0565: **High-dimensional change-point detection using generalized homogeneity metrics**
Presenter: **Shubhadeep Chakraborty**, University of Washington, Seattle, USA, United States
Co-authors: Xianyang Zhang

Change-point detection has been a classical problem in statistics and econometrics. The focus is on the problem of detecting abrupt distributional changes in the data-generating distribution of a sequence of high-dimensional observations, beyond the first two moments. This has remained a substantially less explored problem in the existing literature, especially in the high-dimensional context, compared to detecting changes in the mean or the covariance structure. We develop a nonparametric methodology to (i) detect an unknown number of change-points in an independent sequence of high-dimensional observations and (ii) test for the significance of the estimated change-point locations. Our approach essentially rests upon nonparametric tests for the homogeneity of two high-dimensional distributions. We construct a single change-point location estimator via defining a cumulative sum process in an embedded Hilbert space. As the key theoretical innovation, we rigorously derive its limiting distribution under the high dimension medium sample size (HDMSS) framework. Subsequently, we combine our statistic with the idea of wild binary segmentation to recursively estimate and test for multiple change-point locations. The superior performance of our methodology compared to other existing procedures is illustrated via extensive simulation studies as well as overstock prices data observed during the period of the Great Recession in the United States.

E0803: **Estimating detection in threshold auto-regressive models via dynamic programming**
Presenter: **Ali Shojaie**, University of Washington, United States

Threshold autoregressive (TAR) models are used in many scientific applications, from economics and finance to epidemiology, due to their flexibility. However, existing approaches often assume a fixed number of thresholds as well as a pre-specified threshold variable. Both of these assumptions are somewhat arbitrary and often violated in applications. To address these limitations, we develop a dynamic programming approach to estimate the locations of the thresholds and the autoregressive parameters in high-dimensional TAR models where the threshold variable may need to be selected among a (small) set of candidate variables. We establish the consistency of both the estimated thresholds and the autoregressive parameters for high-dimensional TAR models and illustrate the advantages of the method via simulated and real data examples.

E1474: **Online change-point detection for high-dimensional data**
Presenter: **Jun Li**, Kent State University, United States

Some new procedures are proposed to detect a change for high-dimensional online data. Theoretical properties of the proposed procedures are explored in the high dimensional setting. More precisely, we derive their average run lengths (ARLs) when there is no change point, and expected detection delays (EDDs) when there is a change point. The accuracy of the theoretical results is confirmed by simulation studies. The practical use of the proposed procedures is demonstrated by real data.

EO368 Room Virtual R26 ORDINAL REGRESSION METHODS Chair: Jonathan Schildcrout

E1345: **Transformation models: Pushing the boundaries**
Presenter: **Torsten Hothorn**, University of Zurich, Switzerland

Transformation models have been around for 60 years. The core idea is to transform a distribution of interest, which typically is a rather messing thing, into a nicely behaving distribution prior to analysis. The literature mostly followed two distinct paradigms: Either, a transformation is somehow guestimated without the actual analysis being even aware of such a thing happening or the transformation is treated as a nuisance parameter. Log-transformating count data or "Box-Cox-Transformations" are typical of the former approach and the partial likelihood estimation in Cox models sparked "semi-parametric" inference in similar models. These developments had tremendous success in many disciplines, yet there are limits to what can be done. More recently, it was proposed to actually estimate the necessary transformation explicitly. Thus, the actual model needs to be aware of data transformations and the uncertainty associated with them. While there are some technical issues with such a procedure, it allows many previously hard problems to be solved rather conveniently. Some areas are discussed where fully parameterised transformation models are attractive alternatives to established statistical instruments, such as in regression for discrete, skewed, bounded, or otherwise "difficult" responses, for count regression, in multivariate regression, in penalised regression, and in situations where observations are correlated in some way, most importantly for clustered data.

E1521: **Comparative performance of a semi-parametric generalized linear model in selected analysis settings**
Presenter: **Paul Rathouz**, University of Texas at Austin, United States

A semi-parametric extension of the generalized linear model family has been previously introduced in which the mean model is specified as in any quasi-likelihood formulation, and the response (reference) distribution is fully specified, albeit non-parametrically. The dimension of the non-parametric component in the semi-parametric generalized linear model (SPGLM) is of the same order as the cardinality of the support space of the response. These models have been applied and are often well-suited for settings ranging from ordinal data with finite support to continuous data. After introducing the SPGLM and addressing some important computational advances over the past several years, we will more formally address two important aspects of this modeling framework. First, using available data, we will conduct a comparative analysis between the SPLGM and the popular proportional odds model, focusing on goodness-of-fit and model interpretation. Then, we will show how the fitted SPGLM model can be used to estimate and make inferences on the CDF as a function of covariates, and, time-permitting, how this can be extended to obtain estimates of quantiles as a function of covariates.

E1585: **Longitudinal ordinal models as a general framework for medical outcomes**
Presenter: **Frank Harrell**, Vanderbilt University School of Medicine, United States

Univariate ordinal models can be used to model a wide variety of longitudinal outcomes, using only standard software, through the use of Markov processes. The aim is to show how longitudinal ordinal models unify a wide variety of types of analyses including time to event, recurrent events, continuous responses interrupted by events, and multiple events that are capable of being placed in a hierarchy. Through the use of marginalization over the previous state in an ordinal multi-state transition model, one may obtain virtually any estimand of interest. Both frequentist and Bayesian methods can be used to fit the model and draw inferences.

EO432 Room Virtual R27 THEORY AND APPLICATIONS IN DIMENSION REDUCTION TECHNIQUES Chair: Eliana Christou

E0852: **Sufficient dimension folding in regression via distance covariance for matrix-valued predictors**
Presenter: **Wenhui Sheng**, Marquette University, United States

In modern data, when predictors are matrix/array-valued, building a reasonable model is much more difficult due to the complicated structure. However, dimension folding that reduces the predictor dimensions while keeping its structure is critical in helping to build a useful model. We develop a new sufficient dimension folding method using distance covariance for regression in such a case. The method works efficiently without strict assumptions on the predictors. It is model-free and nonparametric, but neither smoothing techniques nor the selection of tuning parameters is needed. Moreover, it works for both univariate and multivariate response cases. In addition, we propose a new method of local search to estimate the structural dimensions. Simulations and real data analysis support the efficiency and effectiveness of the proposed method.

E0986: **Sufficient dimension folding with categorical predictors**
Presenter: **Qingcong Yuan**, Miami University, United States
Co-authors: Yuanwen Wang, Yuan Xue, Xiangrong Yin

Dimension folding is studied for matrix/array structured predictors with categorical variables. The categorical variable information is incorporated into dimension folding for regression and classification. The concepts of marginal, conditional, and partial folding subspaces are introduced, and their connections to the central folding subspaces are investigated. Estimation methods are proposed to estimate the desired partial folding subspace. An empirical maximal eigenvalue ratio criterion is used to determine the structural dimensions of the associated partial folding subspace. The effectiveness of the proposed methods is evaluated through simulation studies and an application to longitudinal data.

E1325: **Quantile martingale difference divergence for dimension reduction**
Presenter: **Chung Eun Lee**, Baruch College, United States
Co-authors: Haileab Hilafu

The aim is to reduce the dimension of predictors by considering the central quantile subspace. To do so, we use a metric, the quantile martingale difference divergence which measures the quantile dependence of a scalar response variable and a vector of predictors. The proposed dimension-reduction method does not involve user-chosen parameters and does not assume a parametric model, making them simple to implement. Extensive simulations and a real-data illustration are provided to demonstrate the usefulness of the proposed method, which are shown to yield competitive finite-sample performance.

EO499 Room Virtual R28 METHODS FOR CENSORED DATA Chair: Anneleen Verhasselt

E0773: **A nonparametric instrumental approach to endogeneity in competing risks models**
Presenter: **Jad Beyhum**, KU Leuven, Belgium
Co-authors: Jean-Pierre Florens, Ingrid Van Keilegom

Treatment models with duration outcomes, competing risks and random right censoring are discussed. The endogeneity issue is solved using a discrete instrumental variable. We show that the competing risks model generates a non-parametric quantile instrumental regression problem. The cause-specific cumulative incidence, the cause-specific hazard and the subdistribution hazard can be recovered from the regression function. A distinguishing feature of the model is that censoring and competing risks prevent identification at some quantiles. We characterize the set of quantiles for which exact identification is possible and give partial identification results for other quantiles. We outline an estimation procedure and discuss its properties. The finite sample performance of the estimator is evaluated through simulations. We apply the proposed method to the Health Insurance Plan of Greater New York experiment.

E0925: **Semiparametric quantile regression for right censored survival data using two-piece asymmetric distributions**
Presenter: **Worku Biyadgie Ewnetu**, Hasselt University, Belgium
Co-authors: Anneleen Verhasselt, Irene Gijbels

Widely used methods such as Cox proportional hazards, accelerated failure time, and Bennet proportional odds models do not model the quantiles directly, but rather allow the assessment of the influence of the covariates only on the location of the distribution. Quantile regression allows assessing the effects of covariates, not only on a location parameter (such as a mean or median) but also on specific percentiles of the conditional distribution. In recent years, a large family of flexible two-piece asymmetric distributions where the location parameter coincides with a specific quantile of the distribution has been studied. In a conditional (regression) setting the use of such a family of two-piece asymmetric distributions has only been investigated in the complete data case in the literature. We propose a semiparametric procedure to estimate the conditional quantile curves of two-piece asymmetric distributions with right-censored survival data. We use a local likelihood estimation technique in a multiparameter functional form, via which the effect of a covariate on the location, scale, and index of the conditional survival distribution can be assessed. The finite-sample performance of the estimators is investigated via simulations, and the methodology is illustrated with two real data examples.

E1083: **Using frailty models in mathematical epidemiology to reveal how diseases are transmitted**
Presenter: **Steven Abrams**, University of Antwerp and UHasselt, Belgium
Co-authors: Niel Hens, Steffen Unkel, Andreas Wienke

In mathematical epidemiology, frailty models are commonly used for representing individual heterogeneities relevant to the transmission of infectious diseases. More specifically, these frailty models naturally encompass individual differences in susceptibility to infection, infectiousness upon

infection and variation in social activity levels corresponding to specific (effective) contacts relevant to disease spread. Here, we focus on both time-invariant and time-varying frailty models, thereby enabling heterogeneities to evolve over time, for the infection-specific infection hazards. The genesis of these models and how they can be derived from the biological processes underlying disease transmission are demonstrated. Multivariate frailty models including shared and correlated frailties are useful for describing the association between (bivariate) infection times when applied to either right-censored or interval-censored data. Hence, we apply these models to model bivariate current status data thereby unraveling the routes of transmission for these diseases.

EO591 Room Virtual R29 SKETCHING AND RANDOM PROJECTION METHODS FOR MODERN DATA ANALYSIS Chair: Edgar Dobriban

E0239: **Two-sample testing of high-dimensional linear regression coefficients via complementary sketching**
Presenter: **Tengyao Wang**, London School of Economics, United Kingdom

A new method is introduced for two-sample testing of high-dimensional linear regression coefficients without assuming that those coefficients are individually estimable. The procedure works by first projecting the matrices of covariates and response vectors along directions that are complementary in sign in a subset of the coordinates, a process which we call 'complementary sketching'. The resulting projected covariates and responses are aggregated to form two test statistics, which are shown to have essentially optimal asymptotic power under a Gaussian design when the difference between the two regression coefficients is sparse and dense, respectively. Simulations confirm that our methods perform well in a broad class of settings.

E1067: **Sparse sketches with small inversion bias via LEverage Score Sparsified (LESS) embeddings**
Presenter: **Michal Derezinski**, University of Michigan, United States

Sketching is a way of constructing a small representation of a large data matrix by applying to it a random matrix. One of the simplest sketching methods is to use a matrix with i.i.d. Gaussian entries. This approach allows us to draw on the extensive statistical toolkit for working with Gaussian Designs, however, Gaussian sketches are often far too expensive to construct, whereas other more efficient sketching techniques do not retain many of their properties. We propose LEverage Score Sparsified (LESS) Embeddings, a sketching method that preserves the statistical properties of a Gaussian matrix, while retaining the efficiency of state-of-the-art approaches. LESS embeddings are a hybrid of two fast sketching techniques, Leverage Score Sampling and Sparse Embedding Matrices. While they are useful in a range of settings, our primary motivation here is the so-called Inversion Bias, which is a bottleneck when estimating quantities that involve the inverse covariance matrix of the data. Inversion bias arises for instance when using sketching in ordinary least squares, uncertainty quantification and optimal design of experiments. For Gaussian sketches we can easily correct this bias with a simple scaling. By adapting techniques from asymptotic random matrix theory, we show that the same bias correction also works for LESS Embeddings, which improves sketching approximation guarantees for a range of distributed estimation tasks.

E1267: **Adaptive sketching methods**
Presenter: **Mert Pilanci**, Stanford University, United States

Highly efficient randomized solvers are considered for the least-squares problem and a general class of non-linear optimization problems. We show that the projection dimension can be reduced to the effective dimension of the problem, while preserving high-probability convergence guarantees. In this regard, we derive sharp matrix deviation inequalities over ellipsoids for both Gaussian and SRHT embeddings. Specifically, we improve on the constant of a classical Gaussian concentration bound whereas, for SRHT embeddings, our deviation inequality involves a novel technical approach. Leveraging these bounds, we are able to design a practical and adaptive algorithm that does not require knowing the effective dimension beforehand. Our method starts with an initial embedding dimension equal to 1 and, over iterations, increases the embedding dimension up to the effective one at most. Hence, our algorithm improves the state-of-the-art computational complexity for solving regularized least-squares problems and Newton systems in nonlinear optimization. Further, we show numerically that it outperforms standard iterative solvers such as the conjugate gradient method and its pre-conditioned version on several standard machine learning datasets.

EO623 Room Virtual R30 ADVANCES IN HIGH-DIMENSIONAL NETWORK ESTIMATION Chair: Kshitij Khare

E0408: **Bayesian joint inference for multiple directed acyclic graphs**
Presenter: **Xuan Cao**, University of Cincinnati, United States
Co-authors: Kyoungjae Lee

In many applications, data often arise from multiple groups that may share similar characteristics. A joint estimation method that models several groups simultaneously can be more efficient than estimating parameters in each group separately. We focus on unraveling the dependence structures of data based on directed acyclic graphs with a known parent ordering and propose a Bayesian joint inference method for multiple graphs. To encourage similar dependence structures across all groups, a Markov random field prior is adopted. This is the first theoretically supported Bayesian method for joint estimation of multiple directed acyclic graphs. The performance of the proposed method is demonstrated using simulation studies, and it is shown that our joint inference outperforms other competitors. We apply our method to fMRI data for simultaneously inferring multiple brain functional networks.

E1514: **Distributionally robust formulation of the graphical lasso**
Presenter: **Sang-Yun Oh**, University of California, Santa Barbara, United States

Building on a recent framework for distributionally robust optimization, estimation of the inverse covariance matrix is considered for multivariate data. We provide a novel notion of a Wasserstein ambiguity set specifically tailored to this estimation problem, leading to a tractable class of regularized estimators. Special cases include penalized likelihood estimators for Gaussian data, specifically the graphical lasso estimator. As a consequence of this formulation, the radius of the Wasserstein ambiguity set is directly related to the regularization parameter in the estimation problem. Using this relationship, the level of robustness of the estimation procedure corresponds to the level of confidence with which the ambiguity set contains a distribution with the population covariance. Furthermore, the radius can be expressed in closed-form as a function of the ordinary sample covariance matrix. Taking advantage of this finding, we develop a simple algorithm to determine a regularization parameter for the graphical lasso, using only the bootstrapped sample covariance matrices, avoiding repeated evaluation of the graphical lasso algorithm during regularization parameter tuning, for example, with cross-validation. Finally, we numerically study the obtained regularization criterion and analyze the robustness of other automated tuning procedures used in practice.

E1693: **Bayesian inference in high-dimensional mixed frequency regression and VAR models**
Presenter: **Kshitij Khare**, University of Florida, United States

Technological advancements in recent years have enabled organizations to collect, organize, store and analyze very large amounts of data from variables that are available at different temporal frequencies - e.g. monthly, weekly, daily. Such data is commonly referred to as mixed frequency time series data. First, we will focus on mixed frequency regression, where the response variable and the covariates are available at different frequencies (for example, quarterly vs. monthly). We will present a novel Bayesian methodology for (sparse) estimation of the regression coefficients and of the (autoregressive) lag length using a Bayesian nested spike-and-slab framework. Second, we will focus on mixed frequency vector autoregressive (VAR) models, which aim to capture linear temporal interdependencies among multiple time series observed at different frequencies. The issue of over-parameterization in a VAR model becomes more acute in high-dimensional settings where the number of variables is more than or comparable to the sample size. We present a Bayesian approach that achieves parameter reduction through a combination of sparsity and simple structural relationships between appropriate parameters. We will illustrate the efficacy of the proposed approach on simulated data and on real data

from macroeconomics, and establish posterior consistency under high-dimensional scaling where the dimension of the VAR system grows with the sample size.

EO754 Room Virtual R31 RECENT DEVELOPMENT IN COMPLEX FUNCTIONAL DATA Chair: Guanqun Cao

E1731: **Estimation and inference in generalized spatial partially linear varying coefficient models**
Presenter: **Jingru Mu**, Kansas State University, United States

A class of generalized partially linear spatially varying coefficient models for data distributed over complex domains is considered. We approximate the varying coefficient functions and linear coefficients via penalized bivariate splines over triangulation that can handle the complex boundary of the spatial domain. The asymptotic normality of estimated constant coefficients and the consistency of estimated varying coefficients are built up under some regularity conditions. We further propose a model selection approach to identify covariates with constant and varying effects. The performance of the proposed method is evaluated by simulation studies and the crash data in Texas.

E1733: **Additive functional regression for densities as responses**
Presenter: **Kyunghee Han**, University of Illinois at Chicago, United States
Co-authors: Hans-Georg Mueller, Byeong Uk Park

The aim is to propose and investigate additive density regression, a novel additive functional regression model for situations where the responses are random distributions that can be viewed as random densities and the predictors are vectors. Data in the form of samples of densities or distributions are increasingly encountered in statistical analysis and there is a need for flexible regression models that accommodate random densities as responses. Such models are of special interest for multivariate continuous predictors, where unrestricted nonparametric regression approaches are subject to the curse of dimensionality. Additive models can be expected to maintain one-dimensional rates of convergence while permitting a substantial degree of flexibility. This motivates the development of additive regression models for situations where multivariate continuous predictors are coupled with densities as responses. To overcome the problem that distributions do not form a vector space, we utilize a class of transformations that map densities to unrestricted square-integrable functions and then deploy an additive functional regression model to fit the responses in the unrestricted space, finally transforming back to density space. We implement the proposed additive model with an extended version of smooth backfitting and establish the consistency of this approach, including rates of convergence.

E1734: **Functional additive models for optimizing individualized treatment rules**
Presenter: **Hyung Park**, New York University School of Medicine, United States

A novel functional additive model is proposed which is uniquely modified and constrained to model nonlinear interactions between a treatment indicator and a potentially large number of functional and/or scalar pretreatment covariates. The primary motivation for this approach is to optimize individualized treatment rules based on functional data from a randomized clinical trial. We generalize functional additive regression models by incorporating treatment-specific components into additive effect components. A structural constraint is imposed on the treatment-specific components in order to provide a class of additive models with main effects and interaction effects that are orthogonal to each other. If primary interest is in the interaction between treatment and the covariates, as is generally the case when optimizing individualized treatment rules, we can thereby circumvent the need to estimate the main effects of the covariates, obviating the need to specify their form and thus avoiding the issue of model misspecification. The methods are illustrated with data from a depression clinical trial with electroencephalogram functional data as patients pretreatment covariates.

EO794 Room Virtual R35 ROBUST INFERENCE IN CONSTRUCTING DYNAMIC TREATMENT REGIMES Chair: Ashkan Ertefaie

E0369: **Higher-order targeted maximum likelihood estimation and its applications in causal inference**
Presenter: **Mark van der Laan**, University of California Berkeley, United States

Asymptotic linearity and efficiency of targeted maximum likelihood estimators (TMLE) of target features of the data distribution rely on a second-order remainder being asymptotically negligible. The Highly Adaptive Lasso (HAL) provides a general nonparametric MLE that controls this remainder at the desired level, only assuming that the target functional parameters are cadlag and have finite sectional variation norm. However, in finite samples, the second-order remainder can dominate the sampling distribution so that inference based on asymptotic normality would be anti-conservative. We propose a new higher-order (say k-th order) TMLE, generalizing the regular (first-order) TMLE. We prove that it satisfies an exact linear expansion, in terms of efficient influence functions of sequentially defined higher-order fluctuations of the target parameter, with a remainder that is a $k+1$-th order remainder. As a consequence, this k-th order TMLE allows statistical inference only relying on the $k+1$-th order remainder being negligible. We present the theoretical result as well as simulations for the second-order TMLE for nonparametric estimation of the causal quantities, and of the integrated squared density. Its general applications in causal inference for optimal adjustment for baseline and time-dependent confounders is highlighted as well. We also discuss advances in computing higher-order efficient influence functions utilizing HAL.

E0173: **Estimating individualized treatment rules from distributed data collection sites**
Presenter: **Erica Moodie**, McGill University, Canada

Precision medicine is a rapidly expanding area of health research wherein patient-level information is used to inform care via individualized treatment rules (ITR). Identifying the ITR typically requires large data sources, and may necessitate multi-centre collaborations. This may raise concerns about data privacy, management of massive data, and situations where confounding or data collected may differ across sites. We will introduce ITRs and a straightforward, doubly-robust estimation method and discuss approaches to the above-mentioned challenges while producing unbiased estimates of rules that tailor treatment to individual characteristics. The preferred approach is illustrated via an analysis of the impact of antidepressant treatments on BMI.

E0267: **Robust Q-learning**
Presenter: **Robert Strawderman**, University of Rochester, United States

Q-learning is a regression-based approach that is widely used to formalize the development of an optimal dynamic treatment strategy. Finite-dimensional working models are typically used to estimate certain nuisance parameters, and misspecification of these working models can result in residual confounding and/or significant efficiency loss. We propose a robust Q-learning approach that allows estimating such nuisance parameters using data-adaptive techniques. Methodology, asymptotics and simulations will be summarized and highlight the utility of the proposed methods in practice. Data from the "Extending Treatment Effectiveness of Naltrexone" multistage randomized trial will be used to illustrate the proposed methods.

EO374 Room Virtual R36 ARLESTAT: AGEING RISKS AND LONG-TERM IMPACT ON ECONOMY & SOCIETY Chair: Marie Kratz

E1042: **A latent multivariate Gaussian process model for longitudinal ageing survey data**
Presenter: **Juhyun Park**, ENSIIE, France
Co-authors: Evanthia Koukouli, Andrew Titman

The motivation comes from the need of analysing complex survey data coming from the English Longitudinal Study of Ageing, an ongoing biannual longitudinal study which follows up selected individuals aged over 50 years old in England since 2002. This provides valuable resources

to study the ageing process from a multi-dimensional perspective and explore how life domains evolve and interrelate with ageing. However, data on individual traits are mostly gathered through questionnaires and tests resulting in a large collection of non-continuous data which are difficult to model directly. Typically, summary scores or other surrogates are derived as continuous variables or limited latent curve modelling techniques are employed, focusing on single traits. In addition, the longitudinal trend is often not examined as a function of age; instead, the regular data collection time is used as the temporal variable and a single estimate is obtained for the age/age group effect. We develop a latent multivariate Gaussian process modelling framework that allows for simultaneous modelling ordinal longitudinal data, which are sampled irregularly and measure multiple life domains. We propose a latent factor structure for the data at a given time point whilst incorporating individual heterogeneity by assuming a multivariate Gaussian process for individuals' latent domain trajectory. We implement our method based on a version of stochastic EM algorithm and present findings from numerical studies.

E1503: **High-dimensional causal inference that capitalizes on experimental design and computing, illustrated with epigenomics**
Presenter: **Marie-Abele Bind**, Massachusetts General Hospital, United States

In a randomized experiment, no matter how unorthodox the design or the basic statistical analysis, a valid p-value is available. This fact had to be obvious to RA Fisher before 1925, and was recognized for its potential practical utility by Brillinger, Jones, and Tukey (in the context of cloud seeding experiments decades ago in 1978). The practical utility of the idea should be even more evident today because of the widespread availability of high-speed computing and non-traditional statistical methods for data analysis (such as Lasso-based methods), which also rely on high-speed computing to complete in realistic time. Despite the simplicity of the argument, the use of this approach seems to be relatively recondite in current statistical practice. This article attempts to rectify this lacunae through the use of a simple example of a small randomized experiment with a high-dimensional epigenetic outcome.

E0534: **Modeling joint lives within families**
Presenter: **Arthur Charpentier**, UQAM, Canada
Co-authors: Ewen Gallic

Family history is usually seen as a significant factor insurance companies look at when applying for a life insurance policy. Where it is used, family history of cardiovascular diseases, death by cancer, or family history of high blood pressure and diabetes could result in higher premiums or no coverage at all. We use massive (historical) data to study dependencies between life lengths within families. If joint life contracts (between a husband and a wife) have been long studied in actuarial literature, little is known about child and parents dependencies. We illustrate those dependencies using 19th-century family trees in France, and quantify implications in annuities computations. For parents and children, we observe a modest but significant positive association between life lengths. It yields different estimates for remaining life expectancy, present values of annuities, or whole life insurance guarantee, given information about the parents (such as the number of parents alive). A similar but weaker pattern is observed when using information on grandparents.

EO222 Room Virtual R37 STATISTICAL INNOVATIONS IN RESEARCH ON HUMAN BRAIN AND COGNITION Chair: Oystein Sorensen

E0796: **High-dimensional mediation analysis with machine learning**
Presenter: **Martin Lindquist**, Johns Hopkins University, United States

Mediation analysis is used to investigate the role of intermediate variables (mediators) that lie in the path between an exposure and an outcome variable. While significant research has focused on developing methods for assessing the influence of mediators on the exposure-outcome relationship, current approaches do not easily extend to settings where the mediator is high-dimensional. We introduce a novel machine learning-based method for identifying high dimensional mediators. The proposed algorithm iterates between using a machine learning model to map the high-dimensional mediators onto a lower-dimensional space and using the predicted values as input into a standard three-variable mediation model. Hence, the machine learning model is trained to maximize the likelihood of the mediation model. Importantly, the proposed algorithm is agnostic to the machine learning model that is used, providing significant flexibility in the types of situations where it can be used. We illustrate the proposed methodology using data from two functional Magnetic Resonance Imaging (fMRI) studies. Using the proposed approach, we are able to identify brain-based measures that simultaneously encode the exposure variable and correlate with the behavioral outcome.

E1033: **Modeling within-item dependencies in parallel data on test responses and brain activation**
Presenter: **Minjeong Jeon**, UCLA, United States

A joint modeling approach is proposed to analyze dependency in parallel response data. Two types of dependency are defined: higher-level dependency and within-item conditional dependency. While higher-level dependency can be estimated with common latent variable modeling approaches, within-item conditional dependency is a unique kind of information that is often not captured with extant methods, despite its potential to shed new insights into the relationship between the two types of response data. We differentiate three ways of modeling within-item conditional dependency by conditioning on raw values, expected values, or residual values of the response data, which have different implications in terms of response processes. The proposed approach is illustrated with the example of analyzing parallel data on response accuracy and brain activations from a Theory of Mind assessment. The consequence of ignoring within-item conditional dependency is investigated with empirical and simulation studies in comparison to conventional dependency analysis that focuses exclusively on relationships between latent variables.

E1041: **Brain imaging genetics with 40,000 subjects and 3,000 phenotypes**
Presenter: **Lloyd Elliott**, Simon Fraser University, Canada

UK Biobank is a major prospective epidemiological study that is carrying out detailed multimodal brain imaging on 100,000 participants, and includes genetics and ongoing health outcomes. We present a new open resource of GWAS summary statistics, resulting from a greatly expanded set of genetic associations with brain phenotypes, using the 2020 UK Biobank imaging data release of approximately 40,000 subjects, 3,000 phenotypes, and 10 million variants with MAF $\geq 1\%$. We include associations on the X chromosome, and several new classes of image-derived phenotypes (primarily, more fine-grained subcortical volumes, and cortical grey-white intensity contrast). We develop a method to identify clusters of associations across phenotypes (Peaks) and we find 692 replicating clusters of associations, including 12 on the X chromosome. Our novel associations implicate pathways involved in the rare X-linked syndrome STAR (syndactyly, telecanthus and anogenital and renal malformations), Alzheimer's disease and mitochondrial disorders.

EO635 Room Virtual R38 STATISTICAL METHODS AND APPLICATIONS IN SPORTS Chair: Francesco Porro

E0619: **Forced to play too many matches? A deep-learning assessment of crowded schedule**
Presenter: **Marco Delogu**, Universita degli studi di Sassari, Italy
Co-authors: Juan Tena Horrillo, Stefano Cabras

Do important upcoming or recent scheduled tasks affect the current productivity of working teams? How is the impact (if any) modified according to team size or by external conditions faced by workers? We study this issue using association football data where team performance is clearly defined and publicly observed before and after completing different activities (football matches). UEFA Champions League (CL) games affect European domestic league matches in a quasi-random fashion. We estimate this effect using a deep learning model, a novel strategy in this context, that allows controlling for many interacting confounding factors without imposing an ad-hoc parametric specification. This approach is instrumental in estimating performance under what-if situations required in a causal analysis. We find that dispersion of attention and effort to different tournaments significantly worsens domestic performance before/after playing the CL match. However, the size of the impact is higher in

the latter case. Our results also suggest that this distortion is higher for small teams and that, compared to home teams, away teams react more conservatively by increasing their probability of drawing. We discuss the relevance of our results for decision-makers.

E0334: **Players' importance in basketball and the generalized Shapley value**
Presenter: **Rodolfo Metulini**, University of Salerno, Italy
Co-authors: Giorgio Gnecco

The issue of how to measure players' importance in team sports is gaining more and more relevance, mainly because of the advent of new data and advanced technologies, in order to help professional coaches and staff with the final aim of winning the game. Each player's importance has been evaluated, for the first time in basketball, by computing his average marginal contribution to the utility of an ordered subset of players, through a generalized version of the Shapley value. A peculiarity is that the value assumed by the generalized characteristic function of the generalized coalitional game is represented by the probability a certain lineup has to win the game. This probability is estimated by applying a logistic regression model, where the response is represented by the game outcome, and the so-called four Dean's factors are used as explanatory features. By applying the proposed approach to play-by-play data covering fourteen full NBA seasons (from 2004/2005 to 2017/18), we obtain generalized Shapley values for the players of selected teams. It is, in such a way, possible to find those players whose average marginal contribution is higher than expected, by comparing each player's generalized Shapley value with his income.

E0274: **Should you park the bus?**
Presenter: **Tim Swartz**, Simon Fraser University, Canada
Co-authors: Tianyu Guan, Jiguo Cao

The aim is to explore defensive play in soccer. The analysis is predicated on the assumption that the area of the convex hull formed by the players on a team provides a proxy for a defensive style where smaller areas coincide with a greater defensive focus. Whereas the pre-processing of the data is an exercise in data science, the statistical analysis is carried out using simple linear models. The resultant messages are nuanced but suggest that an extremely defensive style is a detrimental strategy in soccer.

EO056 Room Virtual R39 EXPERIMENTS ON NETWORKS Chair: Steven Gilmour

E0328: **Designing experiments on networks: An agricultural field experiment**
Presenter: **Vasiliki Koutra**, King's College London, United Kingdom

The design of experiments provides a formal framework for the collection of data to aid decision making. When such experiments are performed on connected units linked through a network, the resulting design and analysis are more complex; e.g. is the observed response from a given unit due to the direct effect of the treatment applied to that unit, or the result of a network, or viral, effect arising from treatments applied to connected units. We propose a new method of constructing efficient designs with complex blocking structures and network effects, and we illustrate it in an agricultural field experiment. We consider a field trial running at Rothamsted Research, and we compare different optimal designs under different models, including the commonly used designs in such situations. We show that designs that ignore the network structure may lead to poor estimates of the differences between treatment effects.

E1184: **Randomized graph cluster randomization**
Presenter: **Johan Ugander**, Stanford University, United States
Co-authors: Hao Yin

Causal inference under network interference provides a formal framework for measuring network effects using randomized experiments. Experimental designs based on graph cluster randomization (GCR), randomizing units at the level of network clusters, have been shown to greatly reduce variance when measuring network treatment effects, compared to unit-level random assignment. But even so the variance is very often prohibitively large. A randomized version of the GCR design is proposed which is descriptively named randomized graph cluster randomization (RGCR), and which uses a random clustering rather than a single fixed clustering. By considering an ensemble of many different cluster assignments, this design avoids a key problem with GCR where a given unit is sometimes "lucky" or "unlucky" in a given clustering, thereby greatly reducing the variance of network treatment effect estimators in both theory and across extensive simulations.

E1699: **Estimating peer-influence effects under homophily: Randomized treatments and insights**
Presenter: **Edoardo Airoldi**, Fox School of Business, Temple University, United States

Classical approaches to causal inference largely rely on the assumption of lack of interference, according to which the outcome of an individual does not depend on the treatment assigned to others, as well as on many other simplifying assumptions, including the absence of strategic behavior. In many applications, however, such as evaluating the effectiveness of health-related interventions that leverage social structure, assessing the impact of product innovations and ad campaigns on social media platforms, or experimentation at scale in large IT organizations, several common simplifying assumptions are simply untenable. Moreover, being able to quantify aspects of complications, such as the causal effect of interference itself, are often inferential targets of interest, rather than nuisances. We will formalize issues that arise in estimating causal effects when interference can be attributed to a network among the units of analysis, within the potential outcomes framework. We will introduce and discuss several strategies for experimental design in this context centered around a useful role for statistical models. In particular, we wish for certain finite-sample properties of the estimates to hold even if the model catastrophically fails, while we would like to gain efficiency if certain aspects of the model are correct. We will then contrast design-based, model-based and model-assisted approaches to experimental design from a decision theoretic perspective.

EO146 Room Virtual R40 SMALL AREA METHODS Chair: Serena Arima

E0225: **A robust goodness-of-fit test for small area estimation**
Presenter: **Mahmoud Torabi**, University of Manitoba, Canada
Co-authors: Jiming Jiang

A method, originally proposed by R. A. Fisher, is developed into a general procedure, called tailoring, for deriving goodness-of-fit tests that are guaranteed to have a chi-squared asymptotic null distribution. The method has a robustness feature that it works correctly in testing a certain aspect of the model while some other aspects of the model may be misspecified. We apply the method to small area estimation. A connection, and difference, to the existing specification test is discussed. We evaluate the tests' performance theoretically and empirically, and compare it with several existing methods. Our empirical results suggest that the proposed test is more accurate in size, and has either higher or similar power compared to the existing tests. The proposed test is also computationally less demanding than the specification test and other comparing methods. A real-data application is discussed.

E0891: **Empirical best prediction for SAE of categorical variables using finite mixtures of multinomial logistic models**
Presenter: **Maria Giovanna Ranalli**, University of Perugia, Italy
Co-authors: Maria Francesca Marino, Marco Alfo, Nicola Salvati

Many survey variables are categorical in nature and SAE methods based on generalised linear mixed models represent a frequent tool of analysis for prediction. An Empirical Best Prediction (EBP) method is developed for responses in the Exponential Family, based on the use of area-specific, Gaussian, random effects. However, a major drawback of this approach is the computational burden required to derive estimates, compute the EBP and, in particular, provide the corresponding measure of reliability. We introduce a semiparametric EBP for categorical outcomes by extending

a previous approach for univariate responses belonging to the Exponential Family of distributions. This approach leaves the mixing distribution (that is, the distribution of the area-specific random effects) unspecified and estimates it from the observed data via a NonParametric Maximum Likelihood approach. This estimate is known to be a discrete distribution defined over a finite number of locations and leads to the definition of a finite mixture specification. Finite sample properties of the proposal are tested via a simulation study. An application is also provided to data from the Italian Labour Force Survey.

E1193: **Record linkage, measurement error and unit level small area estimation: A Bayesian approach**
Presenter: **Brunero Liseo**, Sapienza Universita di Roma, Italy
Co-authors: Serena Arima, Gauri Datta

Small area estimation is a statistical technique involving the estimation of parameters for small sub-populations, generally used when the sub-population of interest is included in a larger survey. Suppose we have m small areas and the goal is to predict the mean of the variable of interest in each area. Using a set of covariates, we consider a well-known unit-level model. However, as widely discussed in the literature, covariates might be affected by measurement error and ignoring such an error may lead to misleading conclusions in terms of both parameters estimation and the small area mean predictions. We consider the very common situation in which covariates come from a different data file with respect to the data file of the response variable. As a consequence, covariates are affected by two sources of error: measurement error and linking error. We propose a multivariate nested error regression model that accounts for both sources of error. We conduct a noninformative Bayesian inference by assigning an improper prior that we prove to lead to a proper posterior distribution. The model performance is investigated using different simulation scenarios and with a real data application.

CI010 Room K E. Safra (Multi-use 01) BIG DATA AND MACROECONOMICS (VIRTUAL) Chair: Alessia Paccagnini

C0350: **Cointegration in large VARs**
Presenter: **Anna Bykhovskaya**, University of Wisconsin - Madison, United States
Co-authors: Vadim Gorin

Cointegration in vector autoregressive processes (VARs) is analyzed for the cases when both the number of coordinates N and the number of time periods T are large and of the same order. We propose a way to examine a VAR for the presence of cointegration based on a modification of the Johansen likelihood ratio test. The advantage of our procedure over the original Johansen test and its finite sample corrections is that our test does not suffer from over-rejection. This is achieved through novel asymptotic theorems for eigenvalues of matrices in the test statistic in the regime of proportionally growing N and T. Our theoretical findings are supported by Monte Carlo simulations and an empirical illustration.

C0629: **Words speak as loudly as actions: The response of equity prices to macroeconomic announcements**
Presenter: **Chiara Scotti**, Board of Governors of the Federal Reserve System, United States

The impact that macroeconomic news has on equity prices is studied. While the literature has already widely documented the effects of macroeconomic announcements on asset prices, as well as their asymmetric impact during good and bad times, we focus on the reaction to news when the description of the state of the economy-as painted by the Federal Open Market Committee (FOMC) statements-deteriorates. We develop a novel FOMC sentiment index using textual analysis techniques, and find that news has a bigger impact on equity prices during bad times as described by our FOMC sentiment index. This finding is consistent with previous literature, which finds that the stock market's reaction depends on the state of the economy, except that the FOMC's description of the state of the economy is the variable that best explains the variation in the response-more so than the state of the economy itself as measured by real-time indices. Our interpretation is that the reaction of equity prices to news depends on the probability of an increase in interest rates, and the FOMC's description of the state of the economy is one of the best predictors of this probability.

CO354 Room Virtual R25 RECENT ADVANCES IN QUANTILE REGRESSION Chair: Carlos Lamarche

C1124: **Estimating quantile treatment effects for panel data**
Presenter: **Zongwu Cai**, University of Kansas, United States
Co-authors: Ying Fang, Ming Lin, Mingfeng Zhan

Motivated by a previous work that proposed a factor-based model to estimate the average treatment effect with panel data, a quantile treatment effect model for panel data is suggested to characterize the distributional effect of a treatment. We utilize the relationship between conditional cumulative distributional function (CDF) and unconditional CDF to estimate the counterfactual quantile for the treated unit. Also, we derive the asymptotic properties for the proposed quantile treatment effect estimator, together with discussing the choice of control units and covariates. A simulation study is conducted to illustrate our method. Finally, the proposed method is applied to estimate the quantile treatment effects of introducing CSI 300 index futures trading on both the log-return and volatility of the stock market in China.

C1337: **A model-aware approach to quantile regression for cross-sectional data with zero-inflated counts**
Presenter: **Derek Young**, University of Kentucky, United States
Co-authors: Xuan Shi, Carlos Lamarche

Quantile regression for cross-sectional data with (potentially zero-inflated) count responses is addressed using a novel model-aware framework. The model-aware framework transfers the assumed conditional (zero-inflated) discrete distribution to its continuous analogue, thus allowing one to leverage existing quantile regression methods to estimate the conditional quantiles of interest. The proposed approach is followed by fitting a nonlinear function to the estimated conditional quantiles, which yields the estimated quantile effects. One major benefit to this approach is that it mitigates the issue of quantiles crossing, which can occur with existing jittering-based quantile regression methods for count data. Identification, large sample results, and bootstrap routines for the estimated quantile effects are discussed. The small sample performance of the procedure is compared with existing approaches. An analysis of the famous Oregon Health Insurance Experiment data is presented.

C1377: **Quantile regression with an endogenous misclassified binary regressor**
Presenter: **Carlos Lamarche**, University of Kentucky, United States

Misreporting of participation in social programs is common, and it has been increasing in all major surveys. We investigate the estimation of a quantile regression model with endogenous misreporting. We propose a two-step approach and show that the estimator is consistent and asymptotically normal. The identification of the model relies on a parametric first stage and the use of additional measurements including instrumental variables. Simulation studies offer small sample behavior of the proposed estimator in comparison with other approaches. An illustration of the new approach using survey data is considered.

CO176 Room Virtual R32 NEW ADVANCES IN EFFICIENCY AND PRODUCTIVITY ANALYSIS Chair: Helena Veiga

C0888: **The structural and productivity effects of infrastructure provision in developed and developing countries**
Presenter: **Luis Orea**, Univ. Autonoma de Madrid y Universidad de Oviedo, Spain
Co-authors: Inmaculada Alvarez, Luis Serven

Evidence as to the effects of infrastructure provision on structural change and economy-wide productivity is provided using industry-level data for a set of developed and developing countries over the 1995-2010 period. A distinctive feature of our empirical strategy is that it allows the measurement of intra and inter-industry resource reallocations which are directly attributable to the infrastructure provision. In order to achieve the

above objective, we propose a two-level top-down decomposition of aggregate total factor productivity (TFP). While the theoretical decomposition allows measuring inter-industry reallocation effects, our frontier specification permits measuring inter-firm reallocation effects. The production models are estimated considering the endogeneity of capital and labor, infrastructure, and institutional quality. We find a remarkable and positive within-industry productivity change in Europe, but quite small in Latin America and Africa. The productivity growth attributable to reallocation of inputs across industries has improved overall productivity in most countries, especially in Africa. While this better allocation is mainly caused by unexplained factors, communications infrastructure has attenuated this better allocation of resources.

C0840: **Measuring efficiency of Peruvian universities: A stochastic frontier analysis**
Presenter: **Helena Veiga**, Universidad Carlos III de Madrid, Spain
Co-authors: Michael Wiper, Juan Carlos Orosco Gavilan

Estimation of the one-sided error component in stochastic frontier models may erroneously attribute firm characteristics to inefficiency if heterogeneity is unaccounted for. We capture it through exogenous variables which may affect the location, scale, or both parameters of a truncated normal inefficiency distribution. The aim is to measure the efficiency of Peruvian universities via stochastic frontier analysis (SFA). The dataset is a panel of thirty-five Peruvian universities observed in the period 2011-2018. In the stochastic frontier, we have considered the number of students that ended the BA, the number of professors, the budget etc, letting for the inefficiency component the research activity measured by the number of papers published annually per professor, the region where the university is located and the ratio between executed and received budget. Our findings suggest that the inclusion of exogenous variables in the inefficiency distribution is able to capture university heterogeneity and helps to obtain more reliable efficiency scores and rankings.

CO326 Room Virtual R33 NONPARAMETRIC ESTIMATION FOR CAUSAL ANALYSIS Chair: Daniel Henderson

C0174: **A general proposal for model-free difference-in-differences**
Presenter: **Daniel Henderson**, University of Alabama, United States
Co-authors: Stefan Sperlich

A general framework is proposed for model-free difference-in-differences analysis with confounders. Following the natural steps in practice, we start by searching for the preferred data setup, namely the simultaneous selection of confounders and potential data (outcome) transformations. We then offer a test for the credibility of identification assumptions. The treatment effects themselves are estimated in two steps: first, the heterogeneous effects stratified along with the confounders, then second, the average treatment effect(s) for the population(s) of interest. We suggest bootstrap procedures to calculate the standard errors of these estimates and significance tests. We study the asymptotic statistics and the finite sample behavior (via simulations) of our tests and estimators. We address practical issues that arise such as bandwidth selection, incorporating sample weights and dealing with discrete data in both the outcome variable and set of confounders.

C0548: **Abadie's Kappa and weighting estimators of the local average treatment effect**
Presenter: **Tymon Sloczynski**, Brandeis University, United States
Co-authors: Derya Uysal, Jeffrey Wooldridge

The finite sample properties of various weighting estimators of the local average treatment effect (LATE), several of which are based on kappa theorem, are studied. We argue that one of the Abadie estimators, which we show is normalized, is likely to dominate the other in many contexts. A notable exception is in settings with one-sided noncompliance, where certain unnormalized estimators have the advantage of being based on a denominator that is bounded away from zero. We use a simulation study and an empirical application to illustrate our findings. While the performance of kappa-based estimators appears to be context-dependent, the normalized estimator performs very well in all settings under consideration.

CO804 Room Virtual R34 THE ECONOMETRICS OF ASSET PRICING Chair: Fotis Grigoris

C1369: **Confident risk premiums and investments using machine learning uncertainties**
Presenter: **Rohit Allena**, University of Houston, United States

Ex-ante standard errors of risk premium predictions from neural networks (NNs) are derived. Considering standard errors, we identify precise stock-level and portfolio-level return predictions and provide improved investment strategies. The confident high-low strategies that take long-short positions exclusively on stocks with precise risk premia significantly outperform traditional high-low trading portfolios out-of-sample. Optimal mean-variance portfolios incorporating (co)variances of expected return predictions also outperform existing strategies. Economically, time-varying standard errors reflect market uncertainty and spike after financial shocks. In the cross-section, the level and precision of risk premia are correlated, thus NN-based investments deliver more gains in the long positions.

C1462: **Machine learning panel data regressions with heavy-tailed dependent data: Theory and application**
Presenter: **Jonas Striaukas**, UCLouvain & FRS-FNRS, Belgium
Co-authors: Andrii Babii, Eric Ghysels, Ryan Ball

Machine learning regressions are introduced for heavy-tailed dependent panel data potentially sampled at different frequencies. We focus on the sparse-group LASSO regularization. This type of regularization can take advantage of the mixed frequency time series panel data structures and we find that it empirically outperforms the unstructured machine learning methods. We obtain oracle inequalities for the pooled and fixed effects sparse-group LASSO panel data estimators recognizing that financial and economic data can have fat tails. To that end, we leverage a new Fuk-Nagaev concentration inequality for panel data consisting of heavy-tailed tau-mixing processes. We also establish results for valid post-model-selection inference for pooled panel data estimators using HAC-based estimator for the long-run variance.

CO786 Room K2.40 (Hybrid 08) LATEST DEVELOPMENTS IN FINANCIAL ECONOMETRICS Chair: Roxana Halbleib

C0248: **Economic policy uncertainty with Ada-net**
Presenter: **Niels Gillmann**, ifo Institute and TU Dresden, Germany
Co-authors: Ostap Okhrin

A local vector autoregressive model is developed. It allows us to estimate time-varying multivariate models without requiring the model parameters to change at every point in time. The estimation is done by a number of locally homogenous intervals and thereby identifying structural breaks. The local intervals are determined in a sequential testing procedure. This approach is especially suited for short time series since our approach usually results in only a few changes in the coefficients over time. We illustrate the method with simulations and a real data application. Using monthly Economic Policy Uncertainty data for five countries over 20 years, we show that uncertainty is connected across the world in a network. Furthermore, there seem to be three major breaks in our sample. Namely, the Global Financial Crisis, the European sovereign debt crisis and the election of Donald Trump as president of the USA.

C0517: **P-range DCC: A score-driven extension to time-varying correlation models**
Presenter: **Philipp Prange**, Zeppelin University, Germany

A score-driven extension to the well-known dynamic conditional correlation (DCC) model is proposed. The model provides means to capture the time-varying influence from news on correlation dynamics. The recursion to update the news parameter overtime is based on the observations of

past periods. By and large, the model increases the flexibility of DCC-type models whilst maintaining their appealing characteristics for applications with large cross-sections. We demonstrate that the model performs well in a variety of different situations and show that incorporation of the time-varying severity of news enriches the examination of correlation dynamics for a global cross-section of equity markets. More particularly, the article shows that the time-varying parameter can account for significant increases in equity return linkages in response to plunging markets amid the outbreak of COVID-19 in early 2020 and subsequent economic recoveries.

C0868: **Bagged value-at-risk forecast combination**
Presenter: **Ekaterina Kazak**, University of Manchetser, United Kingdom
Co-authors: Roxana Halbleib, Winfried Pohlmeier

Recent developments in financial econometrics literature on joint scoring functions for Value-at-Risk and Expected Shortfall allowed for consistent implementation of statistical tests based on the Model Confidence Set (MCS). MCS is shown to be a great tool for model comparison, both in-sample and out-of-sample. Another branch of literature focused on the superior performance of convex forecast combinations, which often outperform stand-alone forecasting models. The aim is to combine both results and to propose a novel approach to a forecast combination of Value-at-Risk and Expected Shortfall based on the MCS. We exploit the statistical properties of bootstrap aggregation (bagging) and combine competing models based on the bootstrapped probability of the model being in the Confidence Set. The resulting forecast combination allows for a flexible and smooth switch between the underlying models and outperforms the corresponding stand-alone forecasts.

CO042 Room K2.41 (Hybrid 09) TAIL RISK AND DENSITY FORECASTING: NEW TECHNIQUES OR NEW DATA Chair: Massimo Guidolin

C0761: **A flexible predictive density combination model for large financial data in regular and crisis periods**
Presenter: **Francesco Ravazzolo**, Free University of Bozen-Bolzano, Italy
Co-authors: Roberto Casarin, Stefano Grassi, Herman van Dijk

A flexible predictive density combination model is introduced for large financial data sets which allow for dynamic weight learning and model set incompleteness. A dimension reduction allocates the large sets of predictive densities and combination weights to relatively small subsets. Given the representation of the probability model in nonlinear state-space form, efficient simulation-based Bayesian inference is proposed using parallel sequential clustering and filtering, implemented on graphics processing units. The approach is applied to combine predictive densities based on the individual stock returns of daily observations of the S& P500 over a large period which includes the last two (financial) crises. Substantial predictive and economic gains are obtained, in particular, in the tails using Value-at-Risk. Evidence on model set incompleteness and dynamic cluster patterns provide valuable signals for improved modelling and more effective financial policies.

C0844: **Government bonds as market stabilising forces**
Presenter: **Anne-Florence Allard**, University of Bristol, United Kingdom
Co-authors: Hamza Hanbali, Kristien Smedts

The dependence between extreme variations in stocks and government bonds are explored. The goal is to determine whether government bonds can dampen turbulences coming from their local stock market. To this end, copula functions are used and a new model is designed. The proposed model unravels significant differences in stock-bond comovement across countries. These differences are used to identify which government bonds act as a stabilising force and are better at stabilising financial markets.

C0611: **Option-implied network measures of tail contagion and stock return predictability**
Presenter: **Manuela Pedio**, University of Bristol, United Kingdom

The Great Financial Crisis of 2008-2009 has raised the attention of policy-makers and researchers about the interconnectedness among the volatility of the returns of financial assets as a potential source of risk that extends beyond the usual changes in correlations and includes transmission channels that operate through the higher-order co-moments of returns. We investigate whether a newly developed, forward-looking measure of volatility spillover risk based on option implied volatilities shows any predictive power for stock returns. We also compare the predictive performance of this measure with that of the volatility spillover index proposed which is based on realized, backwards-looking volatilities instead. While both measures show evidence of in-sample predictive power, only the option-implied measure is able to produce out-of-sample forecasts that outperform a simple historical mean benchmark.

CC874 Room K0.18 (Hybrid 03) CONTRIBUTIONS IN RISK ANALYSIS (VIRTUAL) Chair: Leopoldo Catania

C0865: **Nonparametric estimator of tail dependence coefficients: Balancing bias and variance**
Presenter: **Maxime Nicolas**, Universite Paris 1 Pantheon-Sorbonne, France
Co-authors: Matthieu Garcin

In risk management, the Tail Dependence Coefficient (TDC) is generally used to measure the dependence between extreme events. Recent work has focused on the non-parametric estimators of the TDC, depending on a threshold that defines which rank corresponds to the tails of a distribution. We present a new method to select the threshold according to a tradeoff between the bias and variance of the estimator. It combines the theoretical Mean Squared Error (MSE) of the estimator and a parametric estimation of the copula linking observations in the tails. We develop several estimation procedures and compare them with other common estimators in a simulation framework. Finally, the method is used to provide risk measurement in a financial dataset.

C1573: **Joint-VaR: A new risk measure for financial markets**
Presenter: **Elisabetta Mensali**, University of Bologna, Italy
Co-authors: Leopoldo Catania, Alessandra Luati

The Joint Value at Risk (JVaR) is defined as the quantile of the conditional distribution of an asset return, given an upper tail event affecting its log-volatility. The purpose of JVaR is to measure financial risk under a volatility stress scenario. A distinguishing feature of the proposed risk measure is that conditioning events are unobserved. The relations with the VaR and the CoVaR, that is the VaR conditional to some event, are made explicit. The properties of JVaR are studied based on a stochastic volatility representation of the underlying process. We prove that JVaR is leverage consistent, i.e. it is an increasing function of the dependence parameter in the stochastic representation. The difference between the JVaR and its reference state, represented by the VaR, provides a natural tool for monitoring risk under volatility distress. An empirical illustration with S&P500 data shows that, during financial crises, accounting for extreme volatility levels is relevant to monitor the evolution of risk.

C1576: **Extreme volatility risk and FX returns**
Presenter: **Josef Kurka**, UTIA AV CR, v.v.i., Czech Republic

While there is a large literature explaining the risk premia on the stock markets, much less is known about the mechanisms generating the foreign exchange (FX) risk premia. We believe that one of the prominent reasons for the failure to explain the currency risk premia is that preferences are modelled as homogeneous across all investors with regard to the risk tastes, and the risk-return relationship is modelled as linear in risk. Volatility is known to be an important indicator of risk and a crucial pricing factor across different assets, and it is also closely connected to the currency risk premia, however, it has also been modelled to have a linear impact across the whole distribution. In contrast, we propose to employ the Extreme Volatility Risk Factor, which should be of the highest relevance to the risk-averse investor cautious mainly about the tails of currency returns distribution, and should be an important step towards explaining the Forward rate bias puzzle. The empirical results suggest that Extreme Volatility

is significantly priced in the cross-section of FX returns.

CC860 Room K0.20 (Hybrid 05) CONTRIBUTIONS IN TIME SERIES ECONOMETRICS (HYBRID) Chair: Tommaso Proietti

C1615: **A sparse Kalman filter: A non-recursive approach**

Presenter: **Jan Bruha**, CNB, Czech Republic

An algorithm is proposed to filter states and shocks in a state-space model under sparsity. Many interesting economic models (such as linearized DSGE models, trend-cyclical VARs, time-varying VARs, dynamic factor models) can be cast into linear state-space models. Under the conventional Kalman filter, which is essentially a recursive OLS algorithm, all shocks are estimated to be non-zero. Sparsity may be beneficial for statistical efficiency and we argue that for some applications, the sparse solution is natural. The sparsity of filtered shocks is achieved by an elastic-net penalty. The algorithm that can be straightforwardly adapted for non-convex penalties or to achieve robustness to outliers

C1679: **The most predictable aspects of time series**

Presenter: **Tommaso Proietti**, University of Roma Tor Vergata, Italy

The focus is on establishing the most predictable aspects of a univariate time series. Assuming the mutual information between past and future as the measure of predictability, we consider classes of transformations depending on a set of basis functions, that are combined linearly. The coefficients are chosen so as to maximize the mutual information between past and future.

Sunday 19.12.2021 08:15 - 09:55 Parallel Session G – CFE-CMStatistics

EO836 Room K0.16 (Hybrid 02) INNOVATIONS IN EXACT AND APPROXIMATE TIME SERIES ANALYSIS Chair: Maryclare Griffin

E1323: **Spectral methods for small sample time series: A complete periodogram approach**
Presenter: **Junho Yang**, Academia Sinica, Taiwan
Co-authors: Sourav Das, Suhasini Subbarao

The periodogram is a widely used tool to analyze second-order stationary time series. An attractive feature of the periodogram is that the expectation of the periodogram is approximately equal to the underlying spectral density of the time series. However, this is only an approximation, and it is well known that the periodogram has a finite sample bias, which can be severe in small samples. We show that the bias arises because of the finite boundary of observation in one of the discrete Fourier transforms which is used in the construction of the periodogram. We show that by using the best linear predictors of the time series over the boundary of observation we can obtain a "complete periodogram" that is an unbiased estimator of the spectral density. In practice, the "complete periodogram" cannot be evaluated as the best linear predictors are unknown. We propose a method for estimating the best linear predictors and prove that the resulting "estimated complete periodogram" has a smaller bias than the regular periodogram. The estimated complete periodogram and a tapered version of it are used to estimate parameters, which can be represented in terms of the integrated spectral density. We prove that the resulting estimators have a smaller bias than their regular periodogram counterparts. The proposed method is illustrated with simulations and real data.

E1449: **Graphical and thresholding local Whittle estimation**
Presenter: **Marie Duker**, Cornell University, United States

The long-run variance matrix and its inverse, the so-called precision matrix, give, respectively, information about correlations and partial correlations between dependent component series of multivariate time series around zero frequency. Non-asymptotic theory is presented for estimation of the long-run variance and precision matrices for high-dimensional Gaussian time series under general assumptions on the dependence structure including long-range dependence. The results for thresholding and penalizing versions of the classical local Whittle estimator ensure consistent estimation in a possibly high-dimensional regime. The highlight is a concentration inequality of the local Whittle estimator for the long-run variance matrix around the true model parameters. In particular, it handles simultaneously the estimation of the memory parameters which enter the underlying model. An application to financial data will also be shown.

E1627: **The elliptical Ornstein-Uhlenbeck process**
Presenter: **Adam Sykulski**, Lancaster University, United Kingdom
Co-authors: Sofia Olhede

The elliptical Ornstein-Uhlenbeck (OU) process is introduced, which is a generalisation of the well-known univariate OU process to bivariate time series. This process maps out elliptical stochastic oscillations over time in the complex plane, which are observed in many applications of coupled bivariate time series. The appeal of the model is that elliptical oscillations are generated using one simple first-order SDE, whereas alternative models require more complicated vectorised or higher-order SDE representations. The second useful feature is that parameter estimation can be performed robustly and quickly in the frequency domain using FFTs of complex-valued data. We determine some properties of the model including the conditions for stationarity, and the geometrical structure of the elliptical oscillations. We demonstrate the utility of the model by measuring periodic and elliptical properties of Earth's polar motion including the Chandler wobble.

E1686: **Graph inference from multivariate time series with long-range dependence**
Presenter: **Irene Gannaz**, INSA Lyon, France

Brain organization, or functional connectivity, is characterized. This can be estimated by the correlation between signals measuring brain activity. Time series have inhomogeneous long-range dependence properties. An estimation procedure in a semi-parametric framework is proposed, based on a Whittle approximation of the wavelet representation. Asymptotic normality is established for the long-range dependence parameters and the long-range correlations. It is used to show that long-range dependence is associated with brain activity. A graphical representation of functional connectivity is inferred by significance tests on the correlations.

EO054 Room K0.18 (Hybrid 03) PROBABILISTIC TIME SERIES FORECASTING Chair: James Taylor

E1367: **Angular combining of forecasts of probability distributions**
Presenter: **James Taylor**, University of Oxford, United Kingdom

When multiple forecasts are available for a probability distribution, forecast combining enables a pragmatic synthesis of the available information to extract the wisdom of the crowd. A linear opinion pool has been widely used, whereby the combining is applied to the probability predictions of the distributional forecasts. However, this has been criticised on theoretical grounds, prompting the combination to be applied to the quantile forecasts of the distributional forecasts. But it has been argued that this will deliver poorer empirical results. We seek an alternative to combining probabilities and combining quantiles. Looking at the distributional forecasts, combining the probability forecasts can be viewed as vertical combining, with quantile forecast combining seen as horizontal combining. Our alternative approach is to allow combining to take place on an angle between the extreme cases of vertical and horizontal combining. The angle can be optimised using a proper scoring rule. We provide empirical illustration using weekly distributional forecasts of COVID-19 mortality for locations in the United States.

E1371: **Forecasting emergency department length of stay and hospital admissions during the pandemic**
Presenter: **Siddharth Arora**, University of Oxford, United Kingdom
Co-authors: James Taylor

The aim is to deploy predictive modelling in an emergency department (ED) to help improve patient outcomes and assist hospitals to make informed interventions for resource allocation/expansion to meet the challenges posed by the pandemic. Firstly, we investigate the impact of the pandemic on ED patient-flow, by focussing on a multitude of outcomes of interest to the service provider, such as attendances, ambulance arrivals, emergency admissions to the hospital, length of stay, and the reason for attendance. Secondly, we forecast the total ED length of stay (LOS), as communicating these estimates can help reduce patient drop-out rates and improve patient satisfaction. Finally, we predict the risk of emergency admission to the hospital, to assist EDs to prioritize patients. For each low-acuity patient, personalized and probabilistic estimates of their LOS and admission risk are provided at their time of registration at the ED. Our forecasting methodology, which is based on machine learning, is adapted to account for the reorganization of ED triage protocol during the pandemic. We envisage the findings of this study could potentially help facilitate patient risk stratification and case management in the ED, which could also have implications for hospital capacity planning.

E1476: **Learning quantile functions for temporal point processes with recurrent neural splines**
Presenter: **Souhaib Ben Taieb**, University of Mons, Belgium

The combination of temporal point process (TPP) models with deep neural networks provides a powerful and flexible framework for modeling continuous-time event data. Neural TPP models are autoregressive models which characterize the distribution of the next arrival time conditional on the observed history. Various representations of this conditional distribution have been proposed in the literature, including parametric forms for the intensity function, the density function, or the cumulative intensity function. Which function to parametrize and how to parametrize it

is an important design choice. We propose a new recurrent neural TPP model which parametrizes the conditional quantile function of the inter-arrival times with a monotonic rational-quadratic spline. While being flexible, our spline-based parameterization has closed-form expressions for multiple useful quantities such as the expectation, the likelihood function, or any quantile. We also derive a closed-form expression for the Continuous Ranked Probability Score (CRPS) which enables efficient model optimization. Finally, we demonstrate that the proposed model achieves competitive performance compared to state-of-the-art neural TPP models on both synthetic and real-world event sequence data.

E1596: **Using L2 and kernel scores to optimise combinations of density forecasts**
Presenter: **Xiaochun Meng**, University of Sussex, United Kingdom
Co-authors: James Taylor

Combining density forecasts has become common practice for various applications. The optimal weights are often obtained by minimising a chosen proper scoring rule, where the log score is most commonly used in the literature. Unfortunately, with the log score, closed-form solutions generally do not exist for the combining weights. We optimise the weights by minimising L2 and kernel scores. We establish the closed-form representations for the optimal weights, and then use them to incorporate a time-varying structure to provide further improvement in forecast accuracy. We use simulated and real data to illustrate our results.

EO595 Room K0.19 (Hybrid 04) ADVANCED STATISTICAL MODELLING Chair: Andreas Mayr

E1372: **STAR Modeling in the battle against pulmonary tuberculosis: Risk areas and associated risk factors**
Presenter: **Bruno de Sousa**, TopAtlantico, Portugal
Co-authors: Carlos Pires, Dulce Gomes, Patricia Filipe, Ana Costa-Veiga, Carla Nunes

Tuberculosis (TB) is still a major global health problem with the World Health Organization estimating that 10 million people fell ill with the disease in 2019, and considering it one of the Top 10 causes of death worldwide (WHO, 2020). Studies that understand the socio-demographic characteristics and time and spatial distribution of the disease are vital to allocating resources in order to improve National TB Programs. The database includes information from all confirmed Pulmonary TB (PTB) cases notified in Continental Portugal. A descriptive analysis of the main risk factors of the disease and the results of two Structured Additive Regression (STAR) models are presented. The former explores possible spatial and temporal correlations in PTB incidence rates in order to identify the regions of increased incidence rates, while the latter provides the profile of the individuals within these regions.

E0542: **Multivariate conditional transformation models**
Presenter: **Thomas Kneib**, University of Goettingen, Germany

Regression models describing the joint distribution of multivariate response variables conditional on covariate information have become an important aspect of contemporary regression analysis. However, a limitation of such models is that they often rely on rather simplistic assumptions, e.g., a constant dependence structure that is not allowed to vary with the covariates or the restriction to linear dependence between the responses only. We propose a general framework for multivariate conditional transformation models that overcomes such limitations and describes the full joint distribution in a tractable and interpretable yet flexible way. Among the particular merits of the framework are that it can be embedded into likelihood-based inference (including results on asymptotic normality) and allows the dependence structure to vary with the covariates. In addition, the framework scales well beyond bivariate response situations.

E1307: **Multivariate conditional transformation models in practice: Conditional reference region estimation**
Presenter: **Oscar Lado-Baleato**, Universidade de Santiago de Compostela, Spain
Co-authors: Carmen Cadarso Suarez, Francisco Gude, Thomas Kneib

When the results of several continuous diagnostic tests are available for the same patient, a multivariate reference region (MVR) is desirable in order to get a clinical interpretation for those results. An MVR, defined as a region that contains 95% of healthy patients results, allows classifying patients into those apparently healthy, and those with some pathology. In diseases diagnosis, MVRs offer a higher specificity and sensitivity than the application of several univariate reference intervals. Although, MVRs are rarely applied in practice because of interpretability difficulties, and Gaussian assumption restriction. Thus, further statistical research is required in order to provide MVRs with higher applicability, and more straightforward interpretability by physicians. Moreover, as diagnostic tests joint distribution change with patients characteristics, irrespectively of the disease status, covariate-adjusted MVRs are desirable in practice. We present a novel formulation for conditional MVRs based on Multivariate Conditional Transformation Models (MCTMs), a brand new multivariate regression framework. The conditional MVRs places no parametric restriction for the response, and continuous covariates non-linear effects might be estimated from the data. MCTMs reference region proved to be reliable with simulated data, and it solved a real problem in diabetes research.

E0921: **Shrinkage of time-varying effects in panel data models**
Presenter: **Helga Wagner**, Johannes Kepler University, Austria

Regression models for panel data with time-varying effects in a Bayesian framework are considered. We implement shrinkage of regression effects as well as the process variances of the effects to distinguish between constant and time-varying effects as well as significant and insignificant effects by appropriate priors distributions. The inference is accomplished via MCMC using ancillarity-sufficiency interweaving which leads to huge improvements in sampling efficiency. The model is applied to analyse panel data on the annual incomes of mothers returning to the job market after maternity leave.

EO404 Room K0.20 (Hybrid 05) RECENT ADVANCES IN FLEXIBLE DIRECTIONAL STATISTICS Chair: Jose Ameijeiras-Alonso

E0802: **Taking advance of the circular manifold for the analysis of circadian gene expression data**
Presenter: **Yolanda Larriba**, University of Valladolid, Spain
Co-authors: Cristina Rueda

Circadian rhythms are genetically encoded by a molecular clock that generates internal timing of approximately 24 hours. At a molecular level, circadian genes display daily rhythmic expression patterns which vary across tissues or species. The underlying circular structure of such molecular rhythms, which can be seen as oscillatory processes, makes the circular space a suitable manifold to formulate and solve key problems in circadian biology. Among others, within the circular space, the sampling temporal order estimation problem is efficiently addressed. The solution to this problem is based both on a circular approach of the well-known principal component analysis method and on the definition of circular order. In addition, it is proposed the use of a novel non-linear parametric regression model capable of adapting to non-sinusoidal shapes, like those observed in the molecular rhythms, and whose parameters are easily interpretable. The results provide accurate genes' peak phase estimates across different tissues and allow a better comprehension of the complex molecular clock network.

E0934: **Nonparametric regression estimation with a circular response and a functional covariable**
Presenter: **Andrea Meilan-Vila**, Carlos III University of Madrid, Spain
Co-authors: Rosa Crujeiras, Mario Francisco-Fernandez

The analysis of a variable of interest that depends on other variables (s) is a typical issue appearing in many practical problems. Regression analysis provides the statistical tools to address this type of problem. This topic has been deeply studied, especially when the variables are of the Euclidean type. However, there are situations where the data present certain kind of complexities, for example, the involved variables are

of circular or functional type, and the classical regression procedures designed for Euclidean data may not be appropriate. In these scenarios, these techniques would have to be conveniently modified to provide useful results. A nonparametric estimator of the circular regression function for models with a circular response and a functional covariate are introduced. Specifically, a Nadaraya– Watson type estimator is proposed and studied. The asymptotic bias and variance of the estimator, as well as, its asymptotic distribution are calculated. Some guidelines for its practical implementation are provided, checking its sample performance through simulations. Finally, the behavior of the estimator is also illustrated with a real data set.

E0988: **Exploring the utility of a circular mode-based model**
Presenter: **Jose Ameijeiras-Alonso**, Universidade de Santiago de Compostela, Spain
Co-authors: Irene Gijbels, Anneleen Verhasselt

The objective is to provide a new mode-based model and to show its applicability in different contexts. With that objective in mind, we will first review some flexible and unimodal parametric models. Secondly, we will introduce a new two-piece four parameters density. Finally, we will see how to extend this family to different contexts. We will include the extension of the proposed model to the regression setting where the response variable is circular.

E0997: **A general approach for nonparametric regression with circular predictors**
Presenter: **Maria Alonso-Pena**, Universidade de Santiago de Compostela, Spain
Co-authors: Irene Gijbels, Rosa Crujeiras

Circular data are observations that can be represented on the circumference of the unit circle, such as angles and directions, and are found in many different fields, for example, biology, meteorology or even psychology. Moreover, circular observations can be found jointly with other variables, and several real-life problems involve the estimation of a regression function when the predictor (or one of the predictors) is of a circular nature. We introduce a broad approach that allows the nonparametric estimation of regression functions with a circular predictor and a general response, which can be either a continuous or a discrete variable. We will present the estimation procedure, which consists of the maximization of the circular kernel weighted log-likelihood, and address the problem of the selection of the smoothing parameter. The finite sample performance of the estimation method will be explored and some real data applications will be presented.

EO382 Room Virtual R20 RECENT ADVANCES IN BAYESIAN COMPUTATION FOR INTRACTABLE SCENARIOS Chair: Matias Quiroz

E0485: **Efficient marginal likelihood estimation by subsampling thermodynamic integration**
Presenter: **Khue-Dung Dang**, University of Melbourne, Australia
Co-authors: Matias Quiroz, Robert Kohn

In Bayesian statistics, the marginal likelihood has a crucial role in model selection. However, it is difficult to compute the marginal likelihood because that requires integrating over all model parameters. Using ideas from thermodynamic integration (TI), the marginal likelihood of a model can be computed via Markov chain Monte Carlo on a series of modified posterior distributions. This method is easy to implement and requires little tuning but can be computationally costly for large data sets. We propose a method to speed up the estimation of the marginal likelihood by combining TI with data subsampling. We apply the new method for modelling binary and time series data sets and show that it is significantly faster than standard TI yet gives similar results.

E0922: **Spectral subsampling MCMC for stationary multivariate time series**
Presenter: **Matias Quiroz**, University of Technology Sydney, Australia
Co-authors: Mattias Villani, Robert Kohn, Robert Salomone

A subsampling Markov chain Monte Carlo approach is proposed to stationary multivariate time series by subsampling periodogram matrix observations in the frequency domain. We also propose a multivariate generalisation of the autoregressive tempered fractionally differentiated moving average model (ARTFIMA) and establish some of its properties. The new model is shown to provide a better fit compared to multivariate autoregressive moving average models for three real-world examples. We demonstrate that spectral subsampling may provide up to two orders of magnitude faster estimation, while retaining MCMC sampling efficiency and accuracy, compared to spectral methods using the full dataset.

E0929: **Bayesian inference for doubly-intractable likelihoods using the block-Poisson estimator**
Presenter: **Yu Yang**, University of New South Wales, Australia
Co-authors: Robert Kohn, Scott Sisson, Matias Quiroz

Many modeling problems in science involve models with unknown normalizing functions, so-called doubly-intractable models in the Bayesian literature. Inference for such models requires the evaluation of likelihood functions that depend on the parameter of interest. We propose an approach that uses a block-Poisson estimator to estimate the doubly-intractable likelihood unbiasedly. The estimator is not necessarily positive, so the pseudo-marginal MCMC algorithm runs on the absolute value of the likelihood estimate with an importance sampling correction that ensures consistent estimates of the posterior mean of any function. We derive practical guidelines on how to tune the hyper-parameters of the estimator to achieve the optimal balance between sampling efficiency and computational cost. The advantage of the method is demonstrated in three examples. The first example concerns the Ising model, a standard example in this literature. The second example concerns modelling a function with a constrained Gaussian process. The third example concerns estimating the parameters in a model for spherical data.

E1078: **Gibbs sampling for finite mixtures with intractable likelihoods**
Presenter: **Thomas Goodwin**, University of Technology, Sydney, Australia
Co-authors: Matias Quiroz, Christian Evenhuis

A finite mixture model is proposed where some of the components may have intractable likelihoods but are easy to simulate from. Such models include multivariate quantile distributions, such as the g-and-k distribution. To estimate the parameters we develop a Gibbs sampler with data augmentation, where the full conditional for the intractable component is sampled via ABC Markov Chain Monte Carlo. To sample the component indicators, we approximate the intractable likelihood via a semiparametric Bayesian synthetic likelihood approach. We demonstrate the model in a flow-cytometry application with highly irregular components.

EO495 Room Virtual R21 STATISTICAL MODELING FOR STOCHASTIC DIFFERENTIAL EQUATIONS Chair: Masayuki Uchida

E1483: **Asymptotic expansion in volatility parametric estimation revisited**
Presenter: **Nakahiro Yoshida**, University of Tokyo, Japan

In the estimation of a volatility parameter from sampled data in a finite time horizon, the asymptotic expansion of the quasi-maximum likelihood estimator and the quasi-Bayesian estimator are revisited. This problem was approached by the author in the 2010s with the martingale expansion. The basic tools are the theory of asymptotic expansion for Skorokhod integrals, the quasi-likelihood analysis, and the order estimate of a polynomial of multiple Wiener integrals.

E0627: **Parameter estimation for misspecified diffusion processes with noisy, nonsynchronous observations**
Presenter: **Teppei Ogihara**, University of Tokyo, Japan

Forecasting variances of stocks and covariances of stock pairs is an important task to control the loss from stock assets for many financial institutions which hold a huge amount of stocks. The study of high-frequency data becomes more important because huge information of high-frequency data

enable us to forecast stock variances and covariances more accurately. However, there are two problems with the statistical analysis of high-frequency data: market microstructure noise and nonsynchronous observations. We study parametric inference under the existence of market microstructure noise and nonsynchronous observations. We study maximum-likelihood-type estimation for parametric diffusion processes with noisy, nonsynchronous observations, assuming that the true model is contained in the parametric family. We further study the case that this assumption is not satisfied. Such a model is called a misspecified model. We will study the asymptotic theory of a maximum-likelihood-type estimator for misspecified models. In this setting, the maximum-likelihood-type estimator cannot attain the optimal convergence rate $n^{-1/4}$ due to the asymptotic bias. We construct a new estimator which attains the optimal rate by using a bias correction, and show the asymptotic mixed normality.

E1322: **Statistics for SPDEs based on discrete observations**
Presenter: **Mathias Trabs**, Karlsruhe Institute of Technology, Germany
Co-authors: Florian Hildebrandt

Stochastic partial differential equations (SPDEs) combine the ability of deterministic PDE models to describe complex mechanisms with the key feature of diffusion models, namely a stochastic signal which evolves within the system. While SPDEs have been intensively studied in stochastic analysis, their statistical theory is still at its beginnings. We study parameter estimation for a parabolic linear stochastic partial differential equation in one space dimension observing the solution field on a discrete grid in a fixed bounded domain. In particular, we discuss central limit theorems for realized quadratic variations based on temporal and spatial increments as well as on double increments in time and space in an infill asymptotic regime in both coordinates.

E0660: **Regularized bridge-type estimation for SDEs**
Presenter: **Alessandro De Gregorio**, University of Rome La Sapienza, Italy

The aim is to introduce an adaptive penalized estimator for identifying the true reduced parametric model under the sparsity assumption. In particular, we deal with the framework where the unpenalized estimator of the structural parameters needs simultaneously multiple rates of convergence (i.e., the so-called mixed-rates asymptotic behavior). We introduce a bridge-type estimator by taking into account penalty functions involving l-norms. We prove that the proposed regularized estimator satisfies the oracle properties. Our approach is useful for the estimation of stochastic differential equations in the parametric sparse setting. More precisely, under the high-frequency observation scheme, we apply our methodology to an ergodic diffusion and introduce a procedure for the selection of the tuning parameters. Furthermore, we will give some hints about possible future developments of this research topic.

EO200 Room Virtual R22 MODEL ASSESSMENT I Chair: Maria Dolores Jimenez-Gamero

E0718: **On the fitting of the distribution of the excess over a confidence level and the adaption for threshold detection**
Presenter: **Daniel Gaigall**, Leibniz University Hannover, Germany
Co-authors: Julian Gerstenberg

The Cramer-von-Mises distance is applied to the distribution of the excess over a confidence level. Asymptotics of related statistics are investigated, and it is seen that the obtained limit distributions differ from the classical ones. For that reason, new bootstrap techniques for approximation purposes are introduced and justified. As an application, the consistency and the asymptotic exactness of a new goodness-of-fit test for the distribution of the excess over a confidence level are deduced. In addition, the results motivate a new confidence interval for the related fitting error. A practice-oriented usage is the determination of appropriate confidence levels for the fitting of the distribution of the excess over a confidence level, where limit results are derived. The adaption for the well-known problem of threshold detection is outlined and illustrated by a real-data example. Simulation studies investigate the quality of the new approximations.

E0996: **On testing independence of count data**
Presenter: **Bojana Milosevic**, University of Belgrade, Serbia
Co-authors: Dana Bucalo Jelic

A novel class of independence tests specially designed for count data is presented. In particular, the class of test statistics is based on the difference between joint and marginal probability generating functions and can be applied to count data of any arbitrary fixed dimension. The limiting properties of test statistics are explored. In order to compare tests with just a few competitors known so far in terms of powers, several resampling procedures have been utilized to approximate null distribution. Their usage is theoretically justified.

E1537: **Profiled deviation subspaces with the application to change structure detection of high-dimensional data**
Presenter: **Jiaqi Huang**, Beijing Normal University, China
Co-authors: Wenbiao Zhao, Xuehu Zhu, Lixing Zhu

The focus is on dimension reduction of high-dimensional data onto the so-called profiled deviation subspaces such that we can equivalently work on lower-dimensional subspaces for detecting change structures within the sequence of data. Consistently estimating the dimensions of the subspaces is studied. Based on these studies, we propose a minimized criterion that is the minimum of component-wise criteria for the orthogonal components being in the subspaces. The estimation consistency of the number of changes and of their locations are verified with the investigation on what divergence rates of the original dimension of the data and the number of changes can ensure the above estimation consistencies. Numerical studies are conducted to examine the finite sample performance of the proposed method and two real data examples are analyzed for illustration.

E0252: **A general Monte Carlo method for goodness-of-fit testing of random vectors**
Presenter: **Feifei Chen**, Beijing Normal University, China
Co-authors: Maria Dolores Jimenez-Gamero, Simos Meintanis, Lixing Zhu

A general and relatively simple method for the construction of goodness-of-fit tests for multivariate distributions is proposed. The method is based on the characterization of probability distributions via their characteristic function. It leads to test criteria that are convenient from the application point of view and consistent against arbitrary deviations from the model under test.

EO264 Room Virtual R23 ADVANCES IN FUNCTIONAL AND MULTIVARIATE DATA ANALYSIS Chair: Yuko Araki

E0770: **A detection test for adjacent hotspot clusters**
Presenter: **Kunihiko Takahashi**, Tokyo Medical and Dental University, Japan
Co-authors: Hideyasu Shimadzu

Several statistical tests have been widely proposed and used in spatial epidemiology to investigate a regional or temporal tendency in the presence of certain diseases, whether the disease risk is relatively high to other surrounding regions or subsequent time periods. The scan statistic is one of the most potent elements of the cluster detection test based on the maximum likelihood ratio for detecting and evaluating spatial and/or temporal disease clusters; examples include Kulldorff's circular scan statistic along with the SaTScan software, and Tango and Takahashi's flexibly shaped scan statistic implemented in the FleXScan software. These scan tests usually deal with a hotspot cluster model assuming a constantly elevated disease rate within a single cluster. However, they can often detect adjacent hotspot clusters exhibiting different disease rates as though they are a single hotspot cluster. This talk proposes a new test procedure that can accurately evaluate these adjacent hotspot clusters as multiple clusters. We present practical examples applying the proposed procedure and compare the results with ones by conventional procedures.

E0974: **Functional dynamic prediction modeling for hourly ambulance demand data**
Presenter: **Toshihiro Misumi**, Yokohama City University, Japan

A dynamic prediction problem for the hourly ambulance demand in Yokohama City, Japan, is considered. We propose a novel functional dynamic prediction modeling based on the functional response regression with both considering long-term trends and seasonal variations. The proposed model enables us to dynamically predict the hourly emergency ambulance demands. We apply the proposed method to the analysis of emergency ambulance demand of Yokohama City from 2008 to 2019. The effectiveness of the proposed method is investigated by Monte Carlo simulations. Our proposed method shows a higher prediction result compared to the existing models without long-term trend and seasonality variations.

E1072: **A greedy approach to detecting change points for functional data**
Presenter: **Jeng-Min Chiou**, Academia Sinica, Taiwan

A new greedy segmentation method is presented to identify multiple change points for a functional data sequence. The estimator's consistency property holds without the common at-most-one-changepoint condition, and it is robust to the relative positions among the change points. Besides, we derive a test statistic based on the estimator and develop an algorithm to determine the number of change points. We will show the asymptotic properties of the algorithm and explore its finite sample performance through a simulation study.

E1425: **Precious asymptotics and Gaussian equivalence with overparameterized likelihood models**
Presenter: **Masaaki Imaizumi**, The University of Tokyo, Japan

A Gaussian equivalence property of a general class of models with likelihoods is studied. The property of high dimensional or over-parameterized models, i.e. learning results are asymptotically equivalent to those when their internal features or variables are Gaussian, has been analyzed in a number of situations. We argue that this equivalence property holds for models that are trained under optimization of likelihood functions. The proof is based on an idea of the theory of U-statistics. The results are verified by numerical experiments.

EO760 Room Virtual R24 DESIGN OF EXPERIMENTS: CONSTRUCTION AND ANALYSIS Chair: Stelios Georgiou

E0957: **Alternatives of second-order response surface designs**
Presenter: **Stelios Georgiou**, RMIT University, Australia

Response surface methodology (RSM) refers to experimental designs for optimizing or developing processes, initially in manufacturing. A new method is presented and an algorithm is implemented that modifies the axial part in a central composite design to achieve a good D-value and efficiency. The new designs are suitable for sequential experimentation. In comparison with known designs in the same class, the new designs are tested and found to have better D-values on a range of factors. With this new approach, efficient orthogonal designs for response surface methodology were generated for a number of parameters that were previously impossible to construct. The new generated designs and their comparison with known designs from the literature are presented in tables for practitioners' use.

E0960: **Missing observations on response surface designs**
Presenter: **Stella Stylianou**, RMIT University, Australia

New minimax loss response surface designs are constructed. These designs are more robust to one missing design point than the original designs. The new designs are compared with the designs in the literature and they are better in terms of loss and number of runs. Moreover, the new designs are better than the original designs in terms of D-efficiency.

E0905: **Discrete choice experiments: An overview on constructing D-optimal and near-optimal choice sets**
Presenter: **Abdulrahman Alamri**, RMIT University, Australia

Discrete choice experiments (DCEs) help to identify the underlying influences on an individual's choice behaviour. With continued developments in DCE methods, DCEs recently are considered the primary data source for decision-makers in various fields, e.g., health resources, marketing, transport, economics, and the list goes on. DCEs are based on stated preference, thus the construction technique of DCE has an obvious effect on the outputs of stated choice. The developments in the field of research nowadays are occurring at a tremendous speed which makes it hard for many to be at the forefront of research and keep up with the state-of-the-art. The question in many practitioners' minds is which techniques perform better (i.e. given small designs with high efficiency) in a given circumstance. To address these concerns, we compile the most efficient techniques for constructing optimal and near-optimal reduced choice sets of main effects only, under the assumption that all alternatives (options) per choice set are equally attractive. The aim is to review and to compare the known theoretical/mathematical constructions of DCEs from known combinatorial and statistical designs in the literature, such as Hadamard matrices, BIBDs, orthogonal arrays, fractional factorial designs. Also, suggesting potential areas for ongoing research in this direction.

E0981: **Enumeration and evaluation of orthogonal three-level designs with small number of runs for definitive screening**
Presenter: **Haralambos Evangelaras**, University of Piraeus / University of Piraeus Research Center, Greece
Co-authors: Viktor Trapouzanlis

A class of three-level designs for definitive screening in the presence of second-order effects has been previously introduced. These designs are called Definitive Screening Designs (DSD), they possess a foldover structure and guarantee that (i) the estimates of main effects are not biased by two-factor interactions, (ii) the two-factor interactions are not completely confounded with each other, and (iii) the quadratic effects are orthogonal to main effects and not completely confounded with two-factor interactions. With respect to the number of runs, these designs require $(2k+1)$ runs to study k quantitative factors since their construction uses a suitable $k \times k$ matrix C, its foldover and a center point. Therefore, each column of the DSD has $(k-1)$ elements equal to $+1$, $(k-1)$ elements equal to -1 and three zeros so, is mean orthogonal. It must be noted, however, that not all $(2k+1) \times k$ such matrices have orthogonal columns. We search for small *nxk* orthogonal three-level designs that possess the attractive properties of a DSD. Complete lists of non-isomorphic designs are given for $n \leq 33$ runs (n odd) and $2 \leq k \leq (n-1)/2$ columns. We further evaluate the constructed designs for their efficiency to estimate the parameters of second-order models, using the popular D-efficiency criterion.

EO120 Room Virtual R25 STATISTICS FOR HIGH-FREQUENCY PRICE AND VOLATILITY MODELS Chair: Markus Bibinger

E0635: **Existence in the inverse Shiryaev problem**
Presenter: **Yoann Potiron**, Keio University, Japan

The inverse first-passage Shiryaev problem is considered, i.e. for $(W_t)_{t\geq 0}$ a standard Brownian motion and any upper boundary continuous function $g : \mathbb{R}^+ \to \mathbb{R}$ satisfying $g(0) \geq 0$, we define $\tau_g^W := \inf\{t \in \mathbb{R}^+ \text{ s.t. } W_t \geq g(t)\}$, and $f_g^W(t)$ its related density. For any target density function of the form $f : \mathbb{R}^+ \to \mathbb{R}^+$ satisfying some smooth assumptions and any arbitrarily big horizon time $T > 0$, we show the existence of a related boundary $g_{f,T} : \mathbb{R}^+ \to [0,T]$, with $g_{f,T}(0) \geq 0$, which satisfies $f_{g_{f,T}}^W(t) = f(t)$ for $0 \leq t \leq T$. As an example, the exponential distribution $f(t) = \lambda \exp(-\lambda t)$ for $\lambda > 0$ satisfies the assumptions. As $g_{f,T}$ is exhibited as a limit boundary of a subsequence of a piecewise linear boundary, we do not obtain any explicit formula for $g_{f,T}$ as a function of f, nor the unicity of the solution. The results are also proved in the symmetrical two-dimensional boundary case.

E0941: **Identifying mixed fractional stable processes from high-frequency data**
Presenter: **Fabian Mies**, RWTH Aachen University, Germany
Co-authors: Mark Podolskij

The linear fractional stable motion generalizes two popular classes of stochastic processes, namely stable Levy processes, on the one hand, and fractional Brownian motion, on the other hand. Hence, it may be regarded as a basic building block for models for high-frequency economic time series. We study a stylized model consisting of a superposition of independent linear fractional stable motions. We construct estimators for all parameters and derive their asymptotic normality in a high-frequency regime. The conditions for consistency turn out to be sharp for two prominent special cases: (i) for Levy processes, i.e. for the estimation of the successive Blumenthal-Getoor indices, and (ii) for the mixed fractional Brownian motion introduced by Cheridito. In the remaining cases, our results reveal an interesting interplay between the Hurst parameter and the index of stability.

E1561: **Local mispricing and microstructural noise: A parametric perspective**
Presenter: **Ilya Archakov**, University of Vienna, Austria
Co-authors: Torben Andersen, Nikolaus Hautsch, Goekhan Cebiroglu

The classic "martingale-plus-noise" model is extended for high-frequency returns to accommodate an error correction mechanism and endogenous pricing errors. It is motivated by (i) novel empirical evidence documenting that microstructure noise exhibits frequently changing patterns of serial dependence which are interwoven with innovations to the efficient price; (ii) building a bridge between high-frequency econometrics and market microstructure models. We identify temporal pricing error correction and noise endogeneity as complementary components driving high-frequency dynamics and inducing two separate regimes, characterized by the sign of the return serial correlation and an implied bias in realized variance estimates. We document frequent fluctuations between these regimes, which can be associated with price discovery in a setting with incomplete information and learning. The model links critical concepts from high-frequency statistics and market microstructure theory, suggesting new avenues for volatility estimation.

E1404: **Joint tests for jumps in asset prices and their spread**
Presenter: **Lars Winkelmann**, Freie Universitaet Berlin, Germany
Co-authors: Wenying Yao

High-frequency econometric methods are developed to test for jumps in the spread of financial asset prices. We derive a coherent inference procedure that detects a jump in the spread only if either of the two asset prices displays a jump. We formalize the test as a sequential procedure in the context of an intersection union test in multiple testing and introduce a new bivariate-jump test for pre-averaged, intra-day returns. We illustrate the practicability of the proposed test procedure by studying the term spread of U.S. government bonds at FOMC announcements.

EO286 Room Virtual R26 SPATIAL EXTREMES Chair: Raphael Huser

E0339: **Max-convolution processes with random shape indicator kernels**
Presenter: **Pavel Krupskiy**, Melbourne University, Australia
Co-authors: Raphael Huser

A new class of models is introduced for spatial data obtained from max-convolution processes based on indicator kernels with random shapes. We study the tail properties of this class of models and show that these are flexible models that can handle complex dependence structures. We discuss estimation methods for these models and apply them to analyze a wind data set.

E0876: **Simulation of spatial hazard events using extremal principal components**
Presenter: **Christian Rohrbeck**, University of Bath, United Kingdom
Co-authors: Daniel Cooley

Hazard event sets, which correspond to a collection of synthetic extreme events, are an important tool for practitioners to analyse and manage future risks. We summarise our method for generating hazard event sets over a set of spatial locations. We start by deriving a set of extremal principal components, with the first components describing the large-scale structure in the extremal dependence structure. The simulation framework is then based on an approximation of the distribution of the extremal principal components. The approach is dimension-reducing in that it distinctly handles the first and remaining extremal principal components. We conclude by discussing some of our recent extensions and a comparison with other existing methods. All aspects are illustrated by applying our method to study extremal behaviour in an environmental setting.

E0896: **Analysis of extremal behaviour using block sums and averages**
Presenter: **Boris Beranger**, University of New South Wales, Australia
Co-authors: Michael Stewart, Scott Sisson

In many environmental applications, it is common that a series of underlying raw observations are condensed into block sums (or averages) and that those summaries are the only source of information available for analysis. For example, raw observations of air pollutant levels are recorded at some weather stations every 5 minutes but only hourly averages are reported. It may be of interest to estimate the exceedance probability of a high threshold, or the quantile corresponding to a very small upper tail probability for the hourly averages. However, similar questions relating to the underlying raw observations might be of greater interest since they directly focus on the extremes of natural phenomena. We are interested in estimating the extremal behaviour of some underlying observations when only block sums and averages are available. We show that some progress can be made when the underlying distribution is heavy-tailed but relatively close to the Gumbel domain of attraction. We establish some theoretical properties of our estimators, and illustrate their performance through numerical experiments and on a pollution dataset.

E1004: **Patterns in spatio-temporal extremes with an application to red sea surface temperature data**
Presenter: **Marco Oesting**, University of Stuttgart, Germany
Co-authors: Raphael Huser

In many applications in environmental sciences, extreme events exhibit a complex spatial-temporal structure that needs to be described in a compressed way. To this end, we propose to investigate not only the duration of the event, but also patterns originating from the temporal course of characteristics describing the event and its spatial structure. Examples include the magnitude of the event or the size of the affected area. In the mathematical framework of regular variation of stochastic processes, we verify the existence of meaningful limit distributions for the patterns of interest and develop non-parametric estimators. These estimators are applied to analyze extreme sea surface temperature in the Red Sea.

EO352 Room Virtual R27 RECENT ADVANCES IN BAYESIAN CAUSAL MEDIATION ANALYSIS Chair: Xinyuan Song

E1174: **Latent multiple mediation analysis with the Bayesian lasso**
Presenter: **Junhao Pan**, Sun Yat-sen University, China
Co-authors: Lijin Zhang

Mediators have played an essential role in helping researchers understand the mechanism through which the predictors affect the outcome variables. The existence of multiple mediators is also very common in behavior research. Traditional approaches for testing the indirect effects include the Sobel test, percentile bootstrap method, and bias-corrected bootstrap method. When handling multiple mediators simultaneously, the traditional approaches for testing the indirect effects, which include the Sobel test and bootstrap method, are prone to inflated Type I error rates and the overfitting problem. To provide a more effective variable selection tool in multiple mediation analysis, mediation models of observed variables were previously integrated with the frequentist Lasso (least absolute shrinkage and selection operator) method. However, this method has two limitations: (1) it does not take the measurement errors of manifest variables into account; (2) it cannot provide uncertainty information about

indirect effects (e.g., interval estimation). The current study extended the Bayesian Lasso method into the framework of latent multiple mediation models to solve the above-mentioned problems. A Monte Carlo simulation study was conducted to compare the proposed method with the traditional Sobel and bootstrap methods. Recommendations and future directions were also provided based on the findings of the simulation study.

E1168: **Bayesian causal mediation analysis with latent mediators and survival outcome**
Presenter: **Xinyuan Song**, Chinese University of Hong Kong, Hong Kong

A joint modeling approach is developed that incorporates latent traits into causal mediation analysis with multiple mediators and a survival outcome. A linear structural equation model is used to characterize the latent mediators with several highly correlated observable surrogates and depicts the relationships among multiple parallel or causally ordered mediators and the exposure. A proportional hazards model is used to derive the path-specific causal effects on the scale of hazard ratio under the counterfactual framework with a set of sequential ignorability assumptions. A Bayesian approach with Markov chain Monte Carlo algorithm is developed to perform efficient estimation of the causal effects. Posterior propriety theory is established for the proportional hazards model with latent variables. Empirical performance of the proposed method is verified through simulation studies. The proposed model is then applied to a study on the Alzheimer's Disease Neuroimaging Initiative dataset to investigate the causal effects of APOE-ε4 allele on the disease progression, either directly or through potential mediators, such as hippocampus atrophy, ventricle expansion, and cognitive impairment.

E1207: **Mediation analysis for mixture Cox proportional hazards cure models**
Presenter: **Xiaoxiao Zhou**, The Chinese University of HongKong, Hong Kong

Mediation analysis aims to decompose a total effect into specific pathways and investigate the underlying causal mechanism. Although existing methods have been developed to conduct mediation analysis in the context of survival models, none of these methods accommodates the existence of a substantial proportion of subjects who never experience the event of interest, even if the follow-up is sufficiently long. We consider mediation analysis for mixture Cox proportional hazards cure models that cope with the cure fraction problem. Path-specific effects on restricted mean survival time and survival probability are assessed by introducing a partially latent group indicator and applying the mediation formula approach in a three-stage mediation framework. A Bayesian approach with P-splines for approximating the baseline hazard function is developed to conduct analysis. The satisfactory performance of the proposed method is verified through simulation studies. An application of the Alzheimer's Disease (AD) Neuroimaging Initiative dataset investigates the causal effects of APOE-4 allele on AD progression.

E1166: **Bayesian causal forest with AFT model: Estimating heterogeneous treatment effects on a survival outcome**
Presenter: **Rongqian Sun**, The Chinese University of Hong Kong, China
Co-authors: Xinyuan Song

Estimating heterogeneous treatment effects has drawn increasing attention in medical studies, considering that patients with divergent features can undergo a different progression of disease even with identical treatment. We consider a joint framework of Bayesian causal forest (BCF) and accelerated failure time (AFT) model to directly capture the possibly heterogeneous treatment effect through two separate Bayesian additive regression trees (BART). The nonparametric BCF structure controls the regularization imposed on treatment effect and flexibly reflects the complex relationship between pre-treatment covariates, treatment indicator, and survival time while requiring no prespecified functional forms. The AFT model is used to derive the conditional average and sample average treatment effect on the scale of log survival time under the potential outcomes framework. Bayesian backfitting Markov chain Monte Carlo algorithm with blocked Gibbs sampler is conducted for estimation of the causal effects. Simulation studies show the satisfactory performance of the proposed method, especially under a small sample size. The proposed model is then applied to a clinical trial that compares two therapies for HIV-infected patients to demonstrate its usage in detecting and visualizing heterogeneous treatment effects.

EO444 Room Virtual R31 NON-REGULAR STATISTICAL MODELING WITH COMPLETE AND INCOMPLETE DATA Chair: Tsung-I Lin

E0337: **Mixtures of multivariate t linear mixed models with missing information**
Presenter: **Tzy-Chy Lin**, Center for Drug Evaluation, Taiwan

Linear mixed-effects (LME) models have been widely used for longitudinal data analysis as they can account for both fixed and random effects, while simultaneously incorporating the variation on both within and between subjects. In clinical trials, some drugs may be more effective in Westerners than in the Orientals. In this situation, such heterogeneity can be modeled by a finite mixture of LME models. The classical modeling approach for random effects and the errors parts are assumed to follow the normal distribution. However, the normal distribution is sensitive to outliers and intolerance of outliers may greatly affect the model estimation and inference. We propose a robust approach called the mixture of multivariate t LME models with missing information. To facilitate the computation and simplify the theoretical derivation, two auxiliary permutation matrices are incorporated into the model to determine the observed and missing components of each observation. We describe a flexible hierarchical representation of the considered model and develop an efficient Expectation-Conditional Maximization Either (ECME) algorithm for carrying out maximum likelihood estimation. Simulation results and real data analysis are provided to illustrate the performance of the proposed methodology.

E0338: **Multivariate-t linear mixed models for longitudinal data with censored and intermittent missing responses**
Presenter: **Wan-Lun Wang**, Feng Chia University, Taiwan
Co-authors: Tsung-I Lin

Multivariate longitudinal data arising in clinical trials and medical studies often exhibit complex features such as censored responses, intermittent missing values, and atypical or outlying observations. The multivariate-t linear mixed model (MtLMM) has been recognized as a powerful tool for the robust modeling of multivariate longitudinal data in the presence of potential outliers or fat-tailed noises. A generalization of MtLMM, called the MtLMM-CM, is presented to properly adjust for censorship due to detection limits of the assay and missingness embodied within multiple outcome variables recorded at irregular occasions. An expectation conditional maximization either (ECME) algorithm is developed to compute parameter estimates using the maximum likelihood (ML) approach. The asymptotic standard errors of the ML estimators of fixed effects are obtained by inverting the empirical information matrix according to Louis method. The proposed methodology is illustrated on a real dataset from HIV-AIDS studies and a simulation study under a variety of scenarios.

E1632: **Bayesian hierarchical functional mixed effects model with shape constrained Gaussian processes**
Presenter: **Jangwon Lee**, Korea University, Korea, South
Co-authors: Taeryon Choi

A Bayesian hierarchical functional mixed-effects model for grouped data observed unequally spaced time is proposed. The method is formulated as a multivariate functional mixed-effects model whose mean part and random part modeled by a Bayesian spectral analysis with and without shape constrained, such as monotone, convex, U-shaped, and multiple-extremes. By assuming the hierarchical structure on the spectral coefficient, the model can capture an overall mean trend as well as group trend and subject-specific trend. For flexible modeling for serial dependence in the temporal data, we assume that the error term is a multivariate Ornstein-Uhlenbeck process. The inference is performed by Markov chain Monte Carlo methods.

E0359: **Extending finite mixtures of t linear mixed-effects models with concomitant covariates**
Presenter: **Yu-Chen Yang**, National Chung Hsing University, Taiwan

Co-authors: Tsung-I Lin, Luis Mauricio Castro, Wan-Lun Wang

The issue of model-based clustering of longitudinal data has attracted increasing attention in the past two decades. Finite mixtures of Student's-*t* linear mixed-effects (FM-tLME) models have been considered for implementing this task, especially when data contain extreme observations. An extended finite mixtures of Student's-t linear mixed-effects (EFM-tLME) model is presented, where the categorical component labels are assumed to be influenced by the observed covariates. Compared with the naive methods assuming the mixing proportions to be fixed but unknown, the proposed EFM-tLME model exploits a logistic function to link the relationship between the prior classification probabilities and the covariates of interest. To carry out maximum likelihood estimation, an alternating expectation conditional maximization (AECM) algorithm is developed under several model reduction schemes. The technique for extracting the information-based standard errors of parameter estimates is also investigated. The proposed method is illustrated using simulation experiments and real data from an AIDS clinical study.

EO366 Room Virtual R38 FUNCTIONAL DATA ANALYSIS AND APPLICATIONS Chair: Marco Stefanucci

E1237: **Stroke rehabilitation assessment by using wrist-worn sensors**

Presenter: **Xi Chen**, Hainan Medical University, China

Stroke is known as a major global health problem, and for stroke survivors, it is key to monitor the recovery levels. However, traditional stroke rehabilitation assessment methods (such as the popular clinical assessment) can be subjective and expensive, and it is also less convenient for patients to visit clinics at a high frequency. To address this issue, based on wearable sensing, we developed two systems that can predict the assessment score in an objective manner. With wrist-worn sensors, accelerometer data were collected from 59 stroke survivors in free-living environments for a duration of 8 weeks, and we aim to map the week-wise accelerometer data (3 days per week) to the assessment score by developing signal processing and predictive model pipeline. To achieve this, we proposed two sets of new features based on the wavelet domain, which can encode the rehabilitation information from both paralysed/non-paralysed sides while suppressing the high-level noises such as irrelevant daily activities. Based on the proposed features, we further employed the longitudinal mixed-effects model with Gaussian process prior (LMGP) to evaluate the patient's recovery levels. Comprehensive experiments were conducted to evaluate our systems on stroke patients, and the results suggested its effectiveness.

E1342: **Curve classification based on feature weighted models with application to medical data**

Presenter: **Chunzheng Cao**, Nanjing University of Information Science and Technology, China

The classification of complex functional data with small inter-class differences and large intra-class differences in the local features is studied. A new type of feature weighted method is proposed. By constructing appropriate weighting functions, it is possible to amplify important features and reduce the interference of bad features on classification. The contribution regards some feature weighted classification by means of F statistics and the weight function based on mean-variance models to improve the accuracy of classification for complex medical data.

E0951: **Evolution outliers in high dimensional functional time series**

Presenter: **Antonio Elias**, Universidad de Malaga, Spain

Co-authors: Antonio Elias, Salvador Pineda, Juan Miguel Morales

Functional depth measures have been a cornerstone to build outlier detection methods for samples of independent and identically distributed curves. We aim to use this concept to extract information of the temporal dependency of Functional Time Series (FTS) that are sets of sample curves indexed in time. More precisely, we focus on High Dimensional Functional Time Series (HDFTS) that terms the situation when several FTS are under analysis. In this scenario, we use functional depth measures to detect evolution outliers that are individual FTS with abnormal time dependency patterns. The performance of our method is tested by simulating HDTS with mixtures of different time dependency structures, namely, time-independent samples, functional dynamic factor models, functional autoregressive models (FAR), functional moving average models (FMA), functional autoregressive and moving average models (FARMA) and functional autoregressive models with seasonality (SFAR). This methodology is motivated by the analysis of data gathered by Energy Smart Metering Infrastructures. They are a multitude of meters that record numerous features such as energy consumption, household circuit voltage, or photo-voltaic energy generation during long time periods at a very high-frequency rate. In this context, we show our HDFTS outlier detection approach with actual smart meters data related to photovoltaic energy generation and circuit voltage records.

E0989: **funcharts: An R package for the real-time monitoring of multivariate functional data**

Presenter: **Christian Capezza**, University of Naples Federico II, Italy

Co-authors: Fabio Centofanti, Antonio Lepore, Alessandra Menafoglio, Biagio Palumbo, Simone Vantini

The R package funcharts provides control charts for multivariate functional data and is based on the functional regression control chart framework introduced in the recent literature. This package is motivated by many modern industrial applications, where data are increasingly available as functional data or profiles. The quality characteristic of interest is often observed as a scalar or a functional quantity and is also affected by other variables, which in turn can be scalar variables or profiles. Functional regression models can be used to take into account the relationship between these variables. The main objective of funcharts is to provide appropriate methods for statistical process monitoring in these cases, with a particular focus on Phase II monitoring. Since in many industrial applications, functional data are observed over a temporal domain and it is important to give indications as soon as possible on possible anomalies, funcharts allows real-time monitoring, i.e., the monitoring of profiles possibly partially observed up to an intermediate domain point. A real-case study on monitoring CO2 emissions from navigation of a Ro-Pax cruise ship is shown to illustrate the practical applicability of the proposed R package.

EO454 Room Virtual R39 MULTIVARIATE ANALYSIS OF COMPLEX DATA Chair: Thomas Verdebout

E0296: **Generalized functional linear models under non-random sampling**

Presenter: **Sophie Dabo**, University of Lille, France

Co-authors: Christelle Judith Agonkoui, Feriel Bouhadjera

A functional binary choice model is explored in a non-random sample design. In other words, a model is considered in which the response is binary, the explanatory variables are functional, and the sample is stratified with respect to the values of the response variable. A dimension reduction of the spaces of the explanatory random functions based on Karhunen-Loeve expansion is used to define a conditional maximum likelihood estimate of the model. Based on this formulation, several asymptotic properties are given. A finite sample study is proposed to compare the proposed method with the ordinary maximum likelihood method, which ignores the nature of the sampling. The potential of the functional model for integrating special non-random features of the sample, which would have been difficult to see otherwise, is outlined.

E0397: **Testing for more positive expectation dependence with application to model comparison**

Presenter: **Julien Trufin**, Universita Libre de Bruxelles, Belgium

Co-authors: Michel Denuit, Julien Trufin, Thomas Verdebout

Modern data science tools are effective to produce predictions that strongly correlate with responses. Model comparison can therefore be based on the strength of dependence between responses and their predictions. Positive expectation dependence turns out to be attractive in that respect. The present talk proposes an effective testing procedure for this dependence concept and applies it to compare two models. A simulation study is performed to evaluate the performances of the proposed testing procedure. Empirical illustrations using insurance loss data demonstrate the

relevance of the approach for model selection in supervised learning. The most positively expectation dependent predictor can then be autocalibrated to obtain its balance-corrected version that appears to be optimal with respect to Bregman, or forecast dominance.

E0643: **Affine-equivariant inference for multivariate location under L_p loss functions**
Presenter: **Davy Paindaveine**, Universite libre de Bruxelles, Belgium
Co-authors: Alexander Duerre

The fundamental problem of estimating the location of a d-variate probability measure under an L_p loss function is considered. The naive estimator, that minimizes the usual empirical L_p risk, has a known asymptotic behaviour but suffers from several deficiencies for $p \neq 2$, the most important one being the lack of equivariance under general affine transformations. We introduce a collection of L_p location estimators that minimize the size of suitable ℓ-dimensional data-based simplices. For $\ell = 1$, these estimators reduce to the naive ones, whereas, for $\ell = d$, they are equivariant under affine transformations. The proposed class contains in particular the celebrated spatial median and Oja median. Under very mild assumptions, we derive an explicit Bahadur representation result for each estimator in the class and establish asymptotic normality. Under a centro-symmetry assumption, we also introduce companion tests for the problem of testing the null hypothesis that the location μ of the underlying probability measure coincides with a given location μ_0. We compute asymptotic powers of these tests under contiguous local alternatives, which reveals that asymptotic relative efficiencies with respect to traditional parametric Gaussian procedures for hypothesis testing coincide with those obtained for point estimation. Monte Carlo exercises confirm our asymptotic results.

E1084: **On the power of Sobolev tests for isotropy under local rotationally symmetric alternatives**
Presenter: **Thomas Verdebout**, Universite Libre de Bruxelles, Belgium
Co-authors: Eduardo Garcia-Portugues, Davy Paindaveine

One of the most classical problems in multivariate statistics is considered, namely, the problem of testing isotropy, or equivalently, the problem of testing uniformity on the unit hypersphere. Rather than restricting to tests that can detect specific types of alternatives only, we consider the broad class of Sobolev tests. While these tests are known to allow for omnibus testing of uniformity, their non-null behavior and consistency rates, unexpectedly, remain largely unexplored. To improve on this, we thoroughly study the local asymptotic powers of Sobolev tests under the most classical alternatives to uniformity, namely, under rotationally symmetric alternatives. We show in particular that the consistency rate of Sobolev tests does not only depend on the coefficients defining these tests but also on the derivatives of the underlying angular function at zero.

EC869 Room Virtual R40 CONTRIBUTIONS IN SPATIAL AND SPATIO-TEMPORAL STATISTICS Chair: Fabio Sigrist

E0735: **Spherical autoregressive change-point detection with applications**
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 of these areas, the detection of structural breaks or regime shifts in the data stream is key. To this end, in the present work, we aim at generalizing the recently introduced SPHAR(p) process by allowing for temporal changes in its functional parameters and variability structure. Our approach, which intrinsically integrates the spatial and temporal dimensions, could give multiscale insights into both the global and local behavior of changes, and its performance will be tested on a real dataset of global surface temperature anomalies.

E1566: **Latent Gaussian model boosting**
Presenter: **Fabio Sigrist**, Lucerne University of Applied Sciences, Switzerland

Latent Gaussian models and boosting are widely used techniques in statistics and machine learning. Tree-boosting shows excellent predictive accuracy on many data sets, but potential drawbacks are that it assumes conditional independence of samples, produces discontinuous predictions for, e.g., spatial data, and it can have difficulty with high-cardinality categorical variables. Latent Gaussian models, such as Gaussian process and grouped random-effects models, are flexible prior models that allow for making probabilistic predictions. However, existing latent Gaussian models usually assume either a zero or a linear prior mean function which can be an unrealistic assumption. We introduce a novel approach that combines boosting and latent Gaussian models in order to remedy the above-mentioned drawbacks and to leverage the advantages of both techniques. We obtain increased predictive accuracy compared to existing approaches in both simulated and real-world data experiments.

E1711: **Clustering spatially resolved genes at a spot level in spatial trascriptomics with SpaRTaCo**
Presenter: **Andrea Sottosanti**, University of Padua, Italy
Co-authors: Davide Risso

Spatial transcriptomics is a modern sequencing technology that allows measuring the activity of thousands of genes in a tissue sample and map where the activity is occurring. The increasing popularity of such advanced technology has grown the interest for the so-called spatially expressed genes, i.e. genes whose expression in a cell affects the expression in the surrounding ones. Comprehending the functions and the interactions of such genes across multiple cells is of great scientific interest because it might lead to a deeper understanding of several complex biological mechanisms. Another relevant aspect in the analysis of tissues is the classification of the cells. Distinguishing, for example, a tumor cell from a stromal or an immune cell is vital. We present SpaRTaCo, a new advanced statistical method for the analysis of spatial transcriptomic experiments. We investigate the properties of the matrix variate distributions to infer the latent block structure at the base of the data, dividing both the genes and the cells into clusters. This procedure, known in statistical literature as co-clustering, allows us to classify the nature of cells and to detect groups of spatially expressed genes only in some specific areas of the tissue sample.

E1520: **A conditional Gaussian process model for ordinal data and its application in predicting herbicidal performance**
Presenter: **Arron Gosnell**, University of Bath, United Kingdom
Co-authors: Evangelos Evangelou

With the proliferation of screening tools for chemical testing, it is now possible to create vast databases of chemicals easily. On the other hand, the development of a rigorous statistical methodology that can be used to analyse these large databases is in its infancy, and further development to facilitate chemical discovery is imperative. Current methods employed to analyse these data fail to incorporate the chemical structure of the tested compound, and as a result, this feature is unaccounted for in the model. We will discuss the Tanimoto similarity as a measure of closeness between chemical compounds and its use within a Gaussian process model. We will demonstrate the application of the proposed model for analysing data from agricultural experiments to assess the herbicidal performance of chemical compounds. The response variable is ordinal, so a proportional odds model is used, with the cumulative probabilities being functions of the Gaussian process. We will show that accounting for correlation results in improved model performance over a simple mixed-effects model and an alternative random forests model. We will discuss the tools used to overcome certain hurdles in developing the model and the use of proper scoring rules to evaluate model performance.

CI026 Room K E. Safra (Multi-use 01) VARIATIONAL INFERENCE FOR BIG MODELS (VIRTUAL) Chair: Michael Smith

C0155: **Fast and accurate variational inference for models with many latent variables**
Presenter: **Michael Smith**, University of Melbourne, Australia
Co-authors: Ruben Loaiza-Maya, David Nott, Peter Danaher

Models with a large number of latent variables are often used to utilize the information in big or complex data, but can be difficult to estimate.

Variational inference methods provide an attractive solution. These methods use an approximation to the posterior density, yet for large latent variable models, existing choices can be inaccurate or slow to calibrate for large latent variable models. We propose a family of tractable variational approximations that are more accurate and faster to calibrate for this case. It combines a parsimonious approximation for the parameter posterior with the exact conditional posterior of the latent variables. We derive a simplified expression for the re-parameterization gradient of the variational lower bound, which is the main ingredient of optimization algorithms used for calibration. The implementation only requires exact or approximate generation from the conditional posterior of the latent variables, rather than the computation of their density. In effect, our method provides a new way to employ Markov chain Monte Carlo (MCMC) within variational inference.

C0156: **Variational inference for cutting feedback in misspecified models**
Presenter: **David Nott**, National University of Singapore, Singapore
Co-authors: Xuejun Yu, Michael Stanley Smith

Bayesian analyses combine information represented by different terms in a joint Bayesian model. When one or more of the terms is misspecified, it can be helpful to restrict the use of information from suspect model components to modify posterior inference. This is called "cutting feedback", and both the specification and computation of the posterior for such cut models is challenging. The definition of cut posterior distributions as solutions to constrained optimization problems is considered, which naturally leads to optimisation-based variational computation methods. The proposed methods are faster than existing Markov chain Monte Carlo (MCMC) approaches for computing cut posterior distributions by an order of magnitude. It is also shown that variational methods allow for the evaluation of computationally intensive conflict checks that can be used to decide whether or not feedback should be cut. The methods are illustrated in a number of simulated and real examples, including an application where recent methodological advances that combine variational inference and MCMC within the variational optimization are used.

C0157: **Streamlined variational inference for random effects models**
Presenter: **Matt P Wand**, University of Technology Sydney, Australia

Variational inference offers fast approximate inference for graphical models arising in computer science and statistics. However, for models containing random effects, direct application of variational inference principles is not sufficient for fast inference due to the sizes of the relevant design matrices. We explain how the notion of matrix algebraic streamlining is crucial for making variational inference practical for models containing very high numbers of random effects. Both nested higher level and crossed random effect structures are discussed.

CO324 Room K0.50 (Hybrid 06) ASSET PRICING AND THE OPTIONS MARKET (IN-PERSON) Chair: Yifan Li

C0686: **The asset durability premium**
Presenter: **Chi-Yang Tsou**, University of Manchester, United Arab Emirates

The aim is to study how the durability of assets affects the cross-section of stock returns. More durable assets incur lower frictionless user costs but are more "expensive", in the sense that they need more down payments making them hard to finance. In recessions, firms become more financially constrained and prefer "cheaper" less durable assets. As a result, the price of less durable assets is less procyclical and therefore less risky than that of durable assets. We provide strong empirical evidence to support this prediction. Among financially constrained stocks, firms with higher asset durability earn average returns about 5% higher than firms with lower asset durability. We develop a general equilibrium model with heterogeneous firms and collateral constraints to quantitatively account for such a positive asset durability premium.

C0946: **Is skewness priced in empirical markets?**
Presenter: **Jiayu Jin**, The University of Manchester, United Kingdom
Co-authors: Kevin Aretz, Yifan Li

Applying the two-stage GMM approach, the physical stochastic volatility model parameters for all single stocks recorded on markets between 1963 and 2020 are estimated. Relying on a newly proposed consistent skewness estimator that specifically targets the physical skewness of an asset's dollar return, existing theories on the pricing of skewness at both short- and long-horizon in stock markets are further tested.

C0705: **The Ross recovery theorem and the term structure of interest rates**
Presenter: **Liangyi Mu**, Queen's Management School, United Kingdom

An omitted risk-free rate condition in Ross recovery estimation is proposed. The term structure of risk-free rates explains the differences between approaches of Ross recovery estimations and the original Ross recovery. A flat term structure of risk-free rates results in a Ross recovered probability distribution identical to the risk-neutral probability distribution in Ross recovery. After considering risk-free rates with a market example, empirical evidence still shows Ross recovered probability distribution is close to the risk-neutral probability distribution. Besides, some challenges with Ross recovery empirical applications are presented. Ross recovery with a short transition time implies a nonnegative matrix root for the transition matrix with a long transition time. Different least squares estimations are not equivalent when there is no unique and accurate fitting with the market state prices. A sparse spot state price surface probably results in a relatively stable pricing kernel in Ross recovery.

C0740: **Option prices and risk-corrected probabilities of a binary event**
Presenter: **Yujing Gong**, London School of Economics, United Kingdom
Co-authors: Arie Gozluklu, Alex Ferreira

Risk-neutral probabilities of the Brexit referendum are estimated using data from both the options and prediction markets. We also provide a risk-corrected measure of these probabilities using both parametric and non-parametric methods. We find evidence that prediction markets reflected the average opinion of the FX options market using reasonable assumptions about preferences and the relative wealth paths in these two states of the world. However, our probabilities are well below the average intention to vote leave from Internet Poll data, which is itself an imperfect measure of the physical probability. By comparing the intention to vote with our estimated probabilities, given the knowledge of the actual outcome, our results show that FX and prediction markets seem to have substantially underestimated the likelihood of Brexit.

CO503 Room Virtual R18 CAUSAL MACHINE LEARNING Chair: Martin Huber

C0205: **Decomposing causal effect heterogeneity under multiple treatment versions**
Presenter: **Michael Knaus**, University St Gallen, Switzerland

A method is developed to decompose treatment effect heterogeneity when the treatment is not homogeneous and can have multiple versions. It disentangles observed aggregated treatment effect heterogeneity into true effect heterogeneity and heterogeneity due to selection into versions. This allows (i) to avoid spurious discovery of heterogeneous effects, (ii) to detect actual hidden heterogeneity in versions, and (iii) to evaluate the underlying version assignment mechanism. We propose a semiparametric method for estimation and statistical inference for the decomposition parameters. Our framework allows for the use of modern machine learning techniques in the estimation of the underlying causal effects. It can be used to conduct simple joint hypothesis tests that consider all treatment versions simultaneously. This alleviates the need for multiple testing procedures when deciding on the aggregation level of the treatment variable in empirical applications. We analyze heterogeneity due to different types of academic or vocational training in the large scale training program for the disadvantaged youth Job Corps. We find that often curricula are not better allocated than random and only specific age and income groups benefit from the actual allocation.

C0220: **Business analytics meets artificial intelligence: Assessing the demand effects of discounts on Swiss train tickets**
Presenter: **Jonas Meier**, University of Bern, Switzerland

Co-authors: Martin Huber, Hannes Wallimann

The demand effects of discounts on train tickets issued by the Swiss Federal Railways, the so-called supersaver tickets, is assessed based on machine learning. Considering a survey-based sample of buyers of supersaver tickets, we investigate which customer- or trip-related characteristics (including the discount rate) predict buying behavior, namely: booking a trip otherwise not realized by train, buying a first- rather than second-class ticket, or rescheduling a trip (e.g. away from rush hours) when being offered a supersaver ticket. Furthermore, we use causal machine learning to assess the impact of the discount rate on rescheduling a trip, which seems relevant in light of capacity constraints at rush hours. Assuming that (i) the discount rate is quasi-random conditional on our rich set of characteristics and (ii) the buying decision increases weakly monotonically in the discount rate, we identify the discount rates effect among always buyers, who would have traveled even without a discount, based on our survey that asks about customer behavior in the absence of discounts. We find that, on average, increasing the discount rate by one percentage point increases the share of rescheduled trips by 0.16 percentage points among always buyers. Investigating effect heterogeneity across observables suggests that the effects are higher for leisure travelers and during peak hours when controlling several other characteristics.

C0546: **Retrieving grouped local average treatment effects via cLasso**

Presenter: **Nicolas Apfel**, University of Regensburg, Germany

Co-authors: Martin Huber, Henrika Langen, Helmut Farbmacher

In the context of an endogenous binary treatment with heterogeneous effects and multiple instruments, we propose a classifier-Lasso (C-Lasso) procedure to identify complier groups with identical local average treatment effects (LATE), in spite of relying on distinct instruments. Our procedure is based on the fact that the LATE needs to be homogeneous for any two or multiple instruments that (i) satisfy the LATE assumptions and (ii) generate identical complier groups in terms of treatment probabilities given the respective instruments. Under the assumption that a relative majority of instruments with identical complier groups satisfies the LATE assumptions, our procedure permits identifying the valid instruments (the exclusion restriction) in a data-driven way. We also provide a simulation study investigating the finite sample properties of our LATE C-Lasso approach and an empirical application investigating the effect of incarceration on recidivism in the US with judge assignments serving as instruments.

C0789: **The impact of MeToo on language at court: A text-based causal inference approach**

Presenter: **Henrika Langen**, University of Fribourg, Switzerland

This study assesses the effect of the MeToo movement on different quantifiers of the 2015-2020 judicial opinions in sexual violence-related cases from 62 U.S. courts. The judicial opinions are vectorized into bag-of-words and tf-idf vectors in order to study their development over time. Further, different indicators quantify to what extent the judges use a language that implicitly shifts some blame from the victim(s) to the perpetrator(s). These indicators measure how the grammatical structure, the sentiment and the context of sentences mentioning the victim(s) and/or perpetrator(s) change over time. The causal effect of the MeToo movement is estimated by means of Difference-in-Differences comparing the development of the language in opinions on sexual violence and other interpersonal crime-related cases as well as by applying panel event study analysis. The results point at a change in the language at court induced by the MeToo movement which materializes with a substantial time lag. Further, the study considers potential effect heterogeneity with respect to the judges' gender and his/her political affiliation. The study combines causal inference with text quantification methods that are commonly used for classification as well as with indicators from the fields of sentiment analysis, word embedding models and grammatical tagging.

CO507 Room Virtual R29 PANEL DATA WITH CROSS-SECTION DEPENDENCE Chair: Yiannis Karavias

C0308: **Identifying dominant units using graphical models in panel time series data**

Presenter: **Jan Ditzen**, Free University of Bozen-Bolzano, Italy

Co-authors: Francesco Ravazzolo

An extension of the graphical model is proposed to estimate the covariance matrix using LASSO estimators. The extension allows the use of time-dependent data and aims to identify dominant units in a network. In detail, we propose to loop through the columns of a data matrix which represents the cross-sectional units. Within each loop we obtain a selection of relevant regressors, which inform about the dependence structure of the data. We carry out a Mont Carlo simulation to show that our estimator correctly identifies the dominant units. We illustrate our method by applying it to a dataset of house prices in England.

C0361: **Structural breaks in interactive effects panels and the stock market reaction to COVID-19**

Presenter: **Yiannis Karavias**, University of Birmingham, United Kingdom

Co-authors: Paresh Kumar Narayan, Joakim Westerlund

Dealing with structural breaks is an essential step in most empirical economic research. This is particularly true in panel data comprised of many cross-sectional units, which are all affected by major events. The COVID-19 pandemic has affected the global economy; however, its impact on stock markets is still unclear. Most markets seem to have recovered while the pandemic is ongoing, suggesting that the relationship between stock returns and COVID-19 has been subject to structural break. We develop a new break detection toolbox that is applicable to different sized panels, easy to implement and robust to general forms of unobserved heterogeneity. The toolbox, which is the first of its kind, includes a structural change test, a break date estimator, and a break date confidence interval. Application to a panel covering 61 countries from January 3 to September 25, 2020, leads to the detection of a structural break that is dated to the first week of April. The effect of COVID-19 is negative before the break and zero thereafter, implying that while markets did react, the reaction was short-lived. A possible explanation is the quantitative easing programs announced by central banks worldwide in the second half of March.

C0366: **Interactive effects panel data models with general factors and regressors**

Presenter: **Joakim Westerlund**, Lund University, Sweden

Co-authors: Liangjun Su, Bin Peng, Yanrong Yang

A model with general regressors and unobservable factors is considered. An estimator based on iterated principal components is proposed, which is shown to be not only asymptotically normal and oracle efficient, but under certain conditions also free of the otherwise so common asymptotic incidental parameters bias. Interestingly, the conditions required to achieve unbiasedness become weaker the stronger the trends in the factors, and if the trending is strong enough, unbiasedness comes at no cost at all. In particular, the approach does not require any knowledge of how many factors there are, or whether they are deterministic or stochastic. The order of integration of the factors is also treated as unknown, as is the order of integration of the regressors, which means that there is no need to pre-test for unit roots, or to decide on which deterministic terms to include in the model.

C0501: **Panel cointegration bounds testing with common factors**

Presenter: **Anindya Banerjee**, University of Birmingham, United Kingdom

Co-authors: Josep Lluis Carrion-i-Silvestre

A panel data unit root statistic is proposed with cross-section dependence driven by unobserved common factors that are approximated by means of the common correlated effects estimation method. The null hypothesis of panel data unit root focuses on the idiosyncratic component, although the statistical inference is conducted using a bounds-testing strategy. Proceeding in this way, the analysis takes into account the fact that the cross-section dependence might be driven by $I(1)$ non-stationary common factors, $I(0)$ common factors, or a mixture of both $I(1)$ and $I(0)$ common

factors and, therefore, extends some existing proposals in the literature.

CO732 Room Virtual R30 NEW METHODS FOR STRUCTURAL VECTOR AUTOREGRESSIONS Chair: Tomasz Wozniak

C1157: **Disentangling Covid-19, economic mobility, and containment policy shocks**
Presenter: **Annika Camehl**, Erasmus University Rotterdam, Netherlands
Co-authors: Malte Rieth

The dynamic impact of Covid-19, economic mobility, and containment policy shocks are studied. We use Bayesian panel structural vector autoregressions with daily data for 44 countries, identified through sign and zero restrictions. Incidence and mobility shocks raise cases and deaths significantly for two months. Restrictive policy shocks lower mobility immediately, cases after one week, and deaths after three weeks. Non-pharmaceutical interventions explain half of the variation in mobility, cases, and deaths worldwide. These flattened the pandemic curve, while deepening the global mobility recession. The policy tradeoff is 1 p.p. less mobility per day for 9% fewer deaths after two months.

C1213: **Singular vector autoregressions**
Presenter: **Eric Eisenstat**, University of Queensland, Australia
Co-authors: Rodney Strachan

Methods are developed for multivariate time series that are assumed to have a strictly singular spectral density. This assumption is widely consistent with economic theory such as in Dynamic Stochastic General Equilibrium (DSGE) models, where the number of variables is almost always greater than the number of structural shocks. Empirically, this assumption guarantees the existence of a finite order VAR representation under mild regularity conditions. However, in this case, there does not exist a unique probability density function with respect to the Lebesgue measure. We overcome this difficulty by defining a density on a compact submanifold with respect to the Hausdorff measure instead. Accordingly, we develop an HMC algorithm that jointly samples model parameters, the VAR lag length, as well as the number of structural shocks in a fully specified Bayesian framework. The effectiveness of the methodology is demonstrated in an extensive Monte Carlo exercise involving a multi-sector DSGE model. Finally, we use the proposed framework to carry out structural analysis on US macroeconomic data in a sample involving COVID shocks.

C1230: **Prior input sensitivities of posterior MCMC inference via infinitesimal perturbation analysis with application to VARs**
Presenter: **Liana Jacobi**, University Melbourne, Australia
Co-authors: Dan Zhu

A key feature of Bayesian inference is the prior (parameter) dependence of posterior distributions that take high-dimensional integrals that typically require numerical solutions. An efficient numerical approach is introduced for input sensitivity analysis of posterior inference via popular Gibbs Markov Chain Monte Carlo (MCMC) simulation methods. We extend Infinitesimal perturbation analysis (IPA) of the simulation path, widely used in the classical simulation context to assess input sensitivities of stochastic dynamic systems, to computational intensive MCMC dependent sampling. We show that IPA derivatives of the posterior statistics from MCMC inference, based on derivatives of parameter draws, have the desirable asymptotic properties of unbiasedness and consistency. We further recommend the use of automatic differentiation to compute these jacobians efficiently. Hence the approach allows for a comprehensive and exact local sensitivity analysis of MCMC output for all input parameters without requiring analytical expressions (likelihood ratio methods, symbolic differentiation) or the re-running of the algorithm (numerical differentiation). We illustrate the use of our method to assess convergence and prior robustness of inference on model parameters and impulse response functions in an application of Bayesian Vector Autoregression analysis with shrinkage priors for US macroeconomic time-series data.

C1273: **Moment condition verification for structural VARs in the Bayesian empirical likelihood framework**
Presenter: **Paul Nguyen**, University of Melbourne, Australia

A method is developed to verify the validity of moment conditions for the identification of structural VARs in the Bayesian Exponentially Tilted Empirical Likelihood (BETEL) framework through the use of slab-and-spike priors. By building upon previous work, we propose a moment condition verification procedure that requires the estimation of only one unified model, allows for a probabilistic interpretation of moment condition validity, and is consistent with the Bayesian model selection guidelines, as used by Chib, Shin and Simoni. To demonstrate the validity of the moment condition verification procedure, we evaluate its performance in simulation by measuring rejection and non-rejection frequencies in a false discovery rate (FDR) controlled environment. We show that across a range of different identification sources, including exclusion restrictions, heteroskedasticity, higher-order moment restrictions, and instrumental variables, the moment condition verification procedure is able to correctly distinguish between valid and invalid moment restrictions. Furthermore, we demonstrate this procedure with an application to a tax policy structural VAR example, evaluating previously used sources of identification such as narrative-based instruments, higher-order moment conditions, and heteroskedasticity.

CO110 Room Virtual R32 ADVANCES IN FINANCIAL ECONOMETRICS Chair: Toshiaki Watanabe

C0376: **Dynamic factor, leverage and realized covariances in multivariate stochastic volatility**
Presenter: **Yasuhiro Omori**, University of Tokyo, Japan
Co-authors: Yuta Yamauchi

In the stochastic volatility models for multivariate daily stock returns, it has been found that the estimates of parameters become unstable as the dimension of returns increases. To solve this problem, we focus on the factor structure of multiple returns and consider two additional sources of information: first, the realized stock index associated with the market factor, and second, the realized covariance matrix calculated from high-frequency data. The proposed dynamic factor model with the leverage effect and realized measures is applied to ten of the top stocks composing the exchange-traded fund linked with the investment return of the S&P500 index, and the model is shown to have a stable advantage in portfolio performance.

C0648: **Realized stochastic volatility models with skew t distributions**
Presenter: **Makoto Takahashi**, Hosei University, Japan
Co-authors: Yasuhiro Omori, Toshiaki Watanabe, Yuta Yamauchi

Predicting volatility and quantiles of financial returns is essential to measure the financial tail risk such as value-at-risk and expected shortfall. There are two important aspects of volatility and quantile forecasts: the distribution of financial returns and the estimation of the volatility. Building on the traditional stochastic volatility model, the realized stochastic volatility model incorporates realized volatility as the precise estimator of the volatility. Using three types of skew-t distributions, the model is extended to capture the well-known characteristics of the return distribution, namely skewness and heavy tails. In addition to the normal and Student's t distributions, included as the special cases of the skew-t distributions, two of them contain the skew-normal, and hence allows more flexible modeling of the return distribution. The Bayesian estimation scheme via a Markov chain Monte Carlo method is developed and applied to major stock indices. The estimation results show that the negative skewness is evident for both indices whereas the heavy tail is largely captured by the realized stochastic volatility, and thus demonstrate that the model with the skew-normal distribution performs well. On the other hand, the prediction results suggest that incorporating both skewness and heavy tail to daily returns is important for volatility and quantile forecasts, especially in a high-volatility period.

C0476: **Stochastic volatility model with time-varying leverage effect**
Presenter: **Jouchi Nakajima**, Bank of Japan, Japan

Co-authors: Toshiaki Watanabe

A stochastic volatility model with a time-varying leverage effect is proposed. The leverage effect, which is captured by a correlation coefficient between innovations to today's return and tomorrow's volatility in a standard stochastic volatility model, is assumed to evolve according to an autoregressive process. An efficient Bayesian method via Markov chain Monte Carlo is developed for the estimation of the proposed model. An empirical analysis using daily stock returns, foreign exchange rate, and cryptocurrency provides evidence that the leverage effect considerably changes over time.

C0484: **Bayesian network analysis for financial risk management**

Presenter: **Mike So**, The Hong Kong University of Science and Technology, Hong Kong

A Bayesian network is a probabilistic graphical model that models conditional dependence (causation) among variables. In most cases, the true underlying structure of a set of variables is unknown. The number of possible structures grows explosively when we have more variables in the networks. We apply a new MCMC sampling scheme for structural learning of Bayesian networks to analyze financial returns data. Time-series properties of the Bayesian networks are investigated to understand the risk evolution in financial markets. We illustrate our approach using stock data in Hong Kong.

CO114 Room Virtual R33 CLIMATE AND ENERGY ECONOMETRICS Chair: Robinson Kruse-Becher

C0335: **Global, hemispheric, and zonal temperature anomalies: How different are they, and what does this mean**

Presenter: **Marc Gronwald**, International Business School Suzhou, China

The empirical properties of three important climate time series are analyzed: global, and hemispheric, and zonal temperature anomalies. First, the motivation is the recent observation that temperature anomalies in the Northern and Southern hemispheres seem to be governed by different stochastic processes. Second, recent theoretical research has shown that insufficiently taking the so-called Polar Amplification into account can lead to sub-optimal climate policies. Among the applied methods, tests for structural breaks and Unit Root structural break tests are considered. The focus is on the timing of structural breaks and the question of deterministic vs stochastic trends. In addition, tests for temporary explosiveness are also applied. Preliminary results suggest that there are spatial differences: while global and Northern hemispheric temperature anomalies are found to be temporary explosive, this does not apply to Southern hemispheric ones. To summarise, the results improve the understanding of empirical properties of important climate time series, and are highly relevant for policy.

C0404: **Rules vs. discretion in cap-and-trade programs: Evidence from the EU emission trading system**

Presenter: **Marina Friedrich**, VU Amsterdam, Netherlands

Co-authors: Sebastien Fries, Michael Pahle, Ottmar Edenhofer

Long-term commitment is crucial for the dynamic efficiency of intertemporal cap-and-trade programs. Discretionary interventions in such programs could destabilize the market and necessitate subsequent corrective interventions that instigate regulatory instability. We provide evidence for this claim from the EU's cap-and-trade program (EU-ETS). We ground our analysis in the theoretical finance literature and apply a mixed-method approach (time-varying coefficient regression, bubble detection, crash-odds modelling). We find that the recent EU-ETS reform triggered market participants into speculation, which likely led to an overreaction that destabilized the market. We discuss how the smokescreen politics behind the reform, which manifested itself in complex rules, was crucial for this outcome. We conclude that rules only ensure long-term commitment when their impact on prices is predictable.

C0954: **A panel SVAR for European climate policy**

Presenter: **Simone Maxand**, Humboldt-Universitaet zu Berlin, Germany

As manifested in international climate agreements, governments worldwide work on the implementation of economically and socially accepted policies to cut down carbon emissions. With a special focus on Europe, structures as the European trading system and individually set carbon prices should provide tools for limiting emissions. A few recent studies empirically address the effectiveness and consequences of implemented carbon taxes. Along with these studies, we follow previous work and evaluate the effect of European carbon taxes on GDP and employment. Model-wise we stay in a similar frame, but apply a newly developed panel SVAR model identified by a combination of non-Gaussianity and narrative restrictions on the structural shocks. In contrast to the original study, we can identify a significant impact of the carbon tax on GDP and employment in the medium term. We additionally present future directions on how this model provides a system view on the effects of environmental taxes in general.

C1439: **A statistical model of the global carbon budget**

Presenter: **Mikkel Bennedsen**, Aarhus University, Denmark

Co-authors: Eric Hillebrand, Siem Jan Koopman

A dynamic statistical model is proposed for the Global Carbon Budget as represented in the annual data set made available by the Global Carbon Project, covering the sample period 1959-2019. The model connects four main objects of interest: atmospheric carbon dioxide (CO2) concentrations, anthropogenic CO2 emissions, the absorption of CO2 by the terrestrial biosphere (land sink), and by the ocean and marine biosphere (ocean sink). The model captures the global carbon budget equation, which states that emissions not absorbed by either land or ocean sinks must remain in the atmosphere and constitute a flow to the stock of atmospheric concentrations. Emissions depend on global economic activity as measured by World Gross Domestic Product while sink activities depend on the level of atmospheric concentrations and the Southern Oscillation Index. We use the model to determine the time series dynamics of atmospheric concentrations, to assess parameter uncertainty, to compute key variables such as the airborne fraction and sink rate, to forecast the Global Carbon Budget components from forecasts of World Gross Domestic Product and Southern Oscillation, and to conduct scenario analysis based on different possible future paths of global economic activity.

CO130 Room Virtual R34 UNCONVENTIONAL MACRO POLICIES AND EXPECTATIONS Chair: Etsuro Shioji

C0513: **Forward guidance as a monetary policy rule**

Presenter: **Takeki Sunakawa**, Hitotsubashi University, Japan

Co-authors: Mitsuru Katagiri

Many central banks implement forward guidance as a state-contingent policy rather than an exogenous policy action in practice. The effects of forward guidance are investigated by formulating it as a systematic part of the monetary policy rule in a non-linear new Keynesian model. It is shown that rule-based forward guidance can significantly mitigate adverse effects on inflation by changing the way of forming expectations about what the central bank can do in a crisis. A quantitative analysis shows that rule-based forward guidance provides new insight about controversial issues including the forward guidance puzzle and the missing deflation puzzle during the Great Recession.

C0831: **Individual trend inflation**

Presenter: **Toshitaka Sekine**, Hitotsubashi University, Japan

Co-authors: Frank Packer, Shunichi Yoneyama

Recent approaches are extended to estimate trend inflation from survey responses of inflation forecasters. In addition to the average inflation forecasts, it tries to estimate the trend inflation of individual forecasters using a noisy information model. While the median of individual trend inflation does not materially differ from that obtained from the aggregate data, it demonstrates much richer analyses are feasible by looking at the

distributions of individual trend inflation. The method is applied to the recent Japanese experiences to see the impacts of the introduction of the 2 percent inflation target and the aggressive monetary easing.

C0641: **The signalling effects of fiscal announcements: Results from supplementary fiscal stimuli**
Presenter: **Hiroshi Morita**, Hosei University, Japan
Co-authors: Leonardo Melosi, Francesco Zanetti

Announcements of fiscal stimulus packages may signal a contractionary view of the government on the future developments in the economy, but the identification of these effects is hindered by the inherent political nature and the large implementation lags of fiscal policy. The supplementary stimulus packages enacted by the Prime Minister Office of Japan over the period 2011-2020 provide an unprecedented set of fiscal stimuli to study the signalling effects of fiscal policy. We show that while fiscal announcements have the standard positive impact on stock prices, the signaling effect of fiscal policy generates a contraction in stock prices when stock market volatility is above the historical average. We develop a simple model of imperfect information that unravels the primary channels of influence for the signalling effects of fiscal announcements. We show that the confidence of agents, the precision of information received by the government and the systematic response of fiscal policy play a chief role in the signalling effects of fiscal policy. The signalling effects of fiscal announcements are contractionary on stock prices when confidence is low, but they turn expansionary when confidence is sufficiently high, and these effects are weakened by the degree of cyclicality in the systematic response of fiscal policy.

C0498: **The pandemic and government bonds: Evidence from volatility smiles in Japan**
Presenter: **Etsuro Shioji**, Hitotsubashi University, Japan

The focus is on how the financial market has reacted to the aggressive fiscal and monetary policies that have been implemented since the outbreak of the COVID-19 in Japan. Even before the pandemic, the country's debt to GDP ratio was well over 200%. The situation has grossly worsened since February 2020. On the other hand, under the Bank of Japan's Yield Curve Control policy, almost half of the Japanese Government Bonds (JGBs hereafter) were in the central bank's hands in the pre-pandemic period. The size of the Bank's balance sheet has further expanded since then. Prices of the JGB futures options are analyzed to measure the policies' influences on private-sector perceptions about the future course of the JGB market. To that end, we derive "volatility smiles" from those option prices on daily basis. We study how the location and the shape of the smile curve have responded to the introduction of the new policy measures.

CO028 Room Virtual R35 TOPICS IN TIME SERIES ECONOMETRICS Chair: Martin Wagner

C1454: **Eigenvalue based monitoring of structural breaks in error correction models**
Presenter: **Leopold Soegner**, Institute for Advanced Studies, Austria
Co-authors: Martin Wagner

Time-series integrated of order one are assumed to be generated by a vector error correction model. Consistent monitoring procedures are developed with the goal to detect structural breaks. Online break-point tests based on the stability of eigenvalues arising in maximum likelihood estimation are obtained. Our focus is on changes in the adjustment coefficients or/and changes in the cointegrating relationships. Breaks where the cointegration rank stays constant as well as breaks where the cointegration rank changes are investigated.

C1712: **A fixed-b cointegration test for cointegrating polynomial regressions**
Presenter: **Sebastian Veldhuis**, University of Klagenfurt, Austria
Co-authors: Martin Wagner

A fixed-b inference is developed for a type of cointegration test for full design cointegrating polynomial regressions, which include, e.g., specifications used in the environmental Kuznets curve literature or Translog-type production and cost functions. Full design refers to a situation that allows reformulating the test statistic as a functional of standard Wiener processes. A detailed simulation study shows that fixed-b inference reduces size distortions significantly at the expense of only modest losses in size-corrected power, a typical feature of fixed-b inference in both stationary as well as unit root and cointegration settings. We illustrate the performance of the test with an application to the environmental Kuznets curve. A supplementary appendix provides critical values for a large number of specifications, a fine grid of bandwidths and several widely-used kernels.

C1713: **Nonparametric cointegration analysis of the environmental Kuznets curve**
Presenter: **Fabian Knorre**, TU Dortmund University, Germany
Co-authors: Martin Wagner

A large and growing literature uses unit root and cointegration techniques to investigate the environmental Kuznets curve (EKC) hypothesis that postulates an inverse U-shaped relationship between the level of economic development and pollution or emissions. Given that economic theory does not, typically, lead to specific functional forms of the EKC, it appears natural to resort to nonparametric estimation. Given the nascent state of the nonparametric cointegration literature, it is provided first a large scale simulation assessment of currently available nonparametric cointegration estimators as well as of tests for the null hypothesis of a specific parametric functional form, e.g., a polynomial relationship, which is the dominant specification in empirical EKC analysis. The considered estimators and tests differ, i.a., with respect to the setting for which asymptotic results are derived, e.g., whether the regressor is allowed to be endogenous or restricted to be strictly exogenous, or whether the errors are allowed to be serially correlated or required to be uncorrelated. The simulation setup is designed to also assess the (finite sample) sensitivity of the estimators and tests with respect to violations of some assumptions. Based upon the simulation findings, we perform nonparametric cointegrating EKC analysis using annual data for CO2 emissions, SO2 emissions and GDP for 18 early industrialized countries over the period 1870-2016.

C1710: **Localized fully modified OLS estimation**
Presenter: **Martin Wagner**, University of Klagenfurt, Austria
Co-authors: Matthias Vetter, Rafael Kawka

An extension of cointegration analysis is considered to a situation where the first differences of the analyzed processes are so-called locally stationary processes rather than stationary processes. This allows us to model long-run relationships between time series whilst allowing for more or less turbulent or persistent periods in the analysis. As is common in the cointegrating regression literature, we allow for regressor endogeneity and error serial correlation, now both time-varying because of our locally-stationary setup. The required functional central limit results for this setting are developed which then allow showing that: First, the OLS estimator is consistent but its limiting distribution is contaminated by second-order bias terms, which differ, of course, from the bias terms arising in the standard context. Second, a localized version of the fully modified OLS estimator for a standard cointegration setting, leads to a zero-mean Gaussian mixture limiting distribution. An important difference to the standard cointegration setting is that fully modified based inference requires something like a HAC-type correction. The theoretical analysis is complemented by a simulation study as well as an empirical application to the forward rate unbiasedness hypothesis.

CO485 Room Virtual R36 ADVANCES IN DURATION ANALYSIS Chair: Ralf Wilke

C0263: **Dependent censoring based on copulas**
Presenter: **Ingrid Van Keilegom**, KU Leuven, Belgium
Co-authors: Claudia Czado

Consider a survival time T that is subject to random right censoring, and suppose that T is stochastically dependent on the censoring time C. We

are interested in the marginal distribution of T. This situation is often encountered in practice. Consider for instance the case where T is the time to death of a patient suffering from a certain disease. Then, the censoring time C is for instance the time until the person leaves the study or the time until he/she dies from another disease. If the reason for leaving the study is related to the health condition of the patient or if he/she dies from a disease that has similar risk factors as the disease of interest, then T and C are likely dependent. We propose a new model that takes this dependence into account. The model is based on a parametric copula for the relationship between T and C, and on parametric marginal distributions for T and C. Unlike most other papers in the literature, we do not assume that the parameter defining the copula function is known. We give sufficient conditions on this parametric copula and marginals under which the bivariate distribution of (T,C) is identified. These sufficient conditions are then checked for a wide range of common copulas and marginal distributions. We also study the estimation of the model, and carry out extensive simulations and the analysis of data on pancreas cancer to illustrate the proposed model and estimation procedure.

C0855: **Dependence evaluation for reliability monitoring data by using the multivariate Farlie-Gumbel-Morgenstern copula**
Presenter: **Shuhei Ota**, Kanagawa University, Japan
Co-authors: Mitsuhiro Kimura

In the research field of reliability engineering, multivariate distributions are frequently used to assess the probability that systems fail dependently. We focus on the multivariate Farlie-Gumbel-Morgenstern (FGM) copula that is one of the multivariate distributions. The problem of estimating the multivariate FGM copula parameters is that the maximum likelihood estimation is not practical because the multivariate FGM copula has many parameters to be estimated, and the parameter constraints are complex. We propose an efficient estimation algorithm for the multivariate FGM copula based on the theory of the inference functions for margins. We show that the estimator given by our estimation algorithm has asymptotic normality as well as its performance determined through simulation studies. Finally, we apply the estimation algorithm to real data analysis of the reliability of ball bearings.

C1071: **A single risk approach to the semiparametric copula competing risks model**
Presenter: **Ming Sum Simon Lo**, United Arab Emirates University, United Arab Emirates
Co-authors: Ralf Wilke

A typical situation in competing risks analysis is that the researcher is only interested in a subset of risks and no distributional information is available about risks that are not of interest. A depending competing risks model is considered with one risk of interest that is known to have a (semi)parametric model, while the model for the other risk is unknown. Identifiability is shown for popular classes of models such as the accelerated failure time model and the semiparametric proportional hazards model. While the semiparametric model requires at least one covariate, the parametric model is identifiable with and without. Different estimation approaches are suggested which are shown to be $\sqrt{n}-$consistent. The applicability is demonstrated with the help of simulations and data examples.

C1580: **Copula based Cox proportional hazards models for dependent censoring**
Presenter: **Negera Wakgari Deresa**, KU Leuven, Belgium
Co-authors: Ingrid Van Keilegom

Most existing copula models for dependent censoring assume that the parameter defining the copula function is known. However, prior knowledge on the dependence is often not available. We propose a novel model that allows the estimation of the copula parameter. The model is based on a parametric copula model for the relation between the survival time T and the censoring time C, where the marginal distributions of T and C follow a semiparametric Cox proportional hazards model and a parametric model, respectively. We show that this model is identified, and propose estimators of the nonparametric cumulative hazard and the finite-dimensional parameters. It is shown that the estimators of the model parameters and the cumulative hazard are consistent and asymptotically normal. We also investigate the performance of the proposed method using finite sample simulations. Finally, we apply our model and estimation procedure to a follicular cell lymphoma data set.

CO643 Room Virtual R37 NETWORK AND REGULARIZATION TECHNIQUES FOR FINANCE Chair: Gabriele Torri

C0395: **Portfolio replication via penalisation with asymmetric deviation measures**
Presenter: **Rosella Giacometti**, VSB-TU Ostrava,Czech Republic, Czech Republic
Co-authors: Gabriele Torri, Sandra Paterlini

Passive strategies, such as the ones implemented by ETFs and ETPs, have gained increasing popularity among investors. In this context, smart beta products promise better performance or lower risk by implementing systematically investing strategies while being cheaper than traditional active strategies. Penalized optimization strategies may offer significant advantages to investors, thanks to their ability to control estimation risk and reduce turnover compared to non-penalized optimal portfolio strategies. In the context of index replication, most literature focuses on the minimization of tracking error measures compared to an index, while imposing constraints on the performances or limiting the number of assets included in the portfolios. We propose here some investment strategies that aim to minimize an asymmetric risk measure related to the expectile, while controlling for the portfolio weight distance from a benchmark composition. Empirical analysis on the S&P100 US, FTSE 100 index and the EUROSTOXX 50 allows to critically discuss the pros and cons of the proposed strategies compared to state-of-art benchmarks

C0937: **Catastrophic and systemic risk in the non-life insurance sector: A micro-structural contagion approach**
Presenter: **Davide Radi**, VSB - Technical University of Ostrava, Italy
Co-authors: Gabriele Torri, Hana Dvorackova

Catastrophic events related to extreme weather and climate change are increasingly common and disrupting. The non-life insurance sector plays a key role in protecting the economic and financial system, and it is therefore of paramount importance to closely monitor its stability. Borrowing from the rich literature of micro-structural interbank contagion, we propose a model to study the stability of a non-life insurance sector in presence of external shocks and random non-diversifiable insurance claims. We show in a simulation study that the stability of such a sector is particularly sensitive to the presence of correlated reinsurance claims (i.e. large undiversifiable shocks, like the ones associated with climate-related catastrophic risk), and that such risk is present even if the reinsurance network is well diversified.

C0973: **Dependency in non-Gaussian settings: The generalized precision matrix and its financial applications**
Presenter: **Gabriele Torri**, University of Bergamo, Italy
Co-authors: Sandra Paterlini, Emanuele Taufer, Rosella Giacometti

Partial correlation networks allow studying the interconnectivity of a financial system. Still, outside the Gaussian framework, they do not allow to fully characterize the interconnectivity structure of random variables. This severely limits its applications in finance where distributions with fat tails and high levels of tail correlation are a better fit for the data. Starting from local dependency measures, we propose a generalization of the precision matrix that describes the interconnectivity structure of multivariate random variables in a single point of the probability space, in a region, or under any conditioning. We use a Gram-Charlier expansion of the density to show how this matrix is related to the traditional precision matrix, we then discuss several parametric cases, focusing on distribution with fat tails, and we illustrate financial applications.

C1762: **Stochastic dominance with uncertain preferences**
Presenter: **Tommaso Lando**, University of Bergamo, Italy

The theory of stochastic dominance provides a model for predicting a decision maker's choice between pairs of uncertain prospects, without having a precise knowledge of her utility function. To improve such a modelization in terms of flexibility, some recent works establish continua of

dominance relations, in which a decision maker's preferences are basically described by a risk aversion or a risk attraction parameter (or both). We study a general theory of stochastic dominance which introduces randomness into such models, by assuming that these parameters, representing preferences, are unknown random variables, to be inferred from data. Under some conditions, this new model makes it possible to estimate the probability of a particular choice between two uncertain prospects and to determine whether it could be expected that the decision-maker will take some decision, given her past behaviour.

CO830 Room K2.40 (Hybrid 08) ECONOMETRIC METHODS AND APPLICATIONS IN TIME SERIES Chair: Shixuan Wang

C0364: **Detecting change points in linear models with homoscedastic or heteroscedastic errors**
Presenter: **Yuqian Zhao**, University of Essex, United Kingdom
Co-authors: Lajos Horvath, Gregory Rice

The problem of detecting change points in the regression parameters of a standard linear model is considered. Motivated by statistics arising from maximally selected likelihood ratio tests, we provide a comprehensive asymptotic theory for weighted functionals of the cumulative sum (CUSUM) process of the residuals, which includes most statistics used for this problem to date. Asymptotic results of this type are then extended to the setting when the model errors and/or covariates exhibit heteroscedasticity. These theoretical results illuminate how to adapt standard change point test statistics to this situation. Such adaptations are studied in a simulation study along with a method based on a classical approximation to improve the finite sample performance of these tests, which show that they work well in practice to detect multiple change points in the linear model parameters, and control the Type-I testing error in the presence of heteroscedasticity. The proposed methods are illustrated with applications to financial sentiment analysis, and to measure for changes in the relationship between COVID-19 infections and deaths in the United Kingdom.

C0875: **A good basis for projections in functional data analysis: Markowitz portfolio optimization**
Presenter: **Shanglin Lu**, Renmin University of China, China
Co-authors: Lajos Horvath, Zhenya Liu

A new basis for projections in functional data analysis is proposed and used to solve the minimum-variance portfolio optimization for large dimensional assets. We measure the large dimensional asset returns in the cross-section as Hilbert-space-valued random function. The proposed basis is formed by the eigenfunctions belonging to the d-largest eigenvalues of a bivariate function, which combines the covariance function and the mean function proportionally. A comprehensive simulation study shows better performance in terms of shrinking error when using the d-dimensional projections in dimension reduction. In the empirical application to the U.S. industry portfolios, Fama and French portfolios, and S&P500 constituent stocks, the optimized portfolio also give better out-of-sample performance by employing the proposed basis.

C0429: **When global warming started**
Presenter: **Zhenya Liu**, Renmin University of China and Aix-Marseille University, France
Co-authors: Xuyuan Han

Based on the longest historical temperature data set, Met Office Hadley Centre Central England Temperature Data, we use the quick testing technique and find the global warming started at 1873-1875, the mean of the start time of global warming is 1874, and the mean of temperature is increased by 0.29 C after 1874. The findings enhance our understanding of the global temperature only reliably go back as far as 1880. In the comparison study with previous work, we find that our result put more weight on the probability of false alarming.

C0731: **Detect the stock price trend in high-frequency data: A stochastic disorder approach**
Presenter: **Yaosong Zhan**, Renmin University of China, China

Identifying the trend of a stock price and its changes is critical for the trend following strategy. We propose a model for detecting trend changes in high-frequency data based on the stochastic disorder theory. We use geometric Brownian motion to describe the stock price, and the compound Poisson process is used to solve the jumps that occur in the high-frequency data. Through solving an optimal stopping time problem, we propose a rule for identifying trend changes. The numerical simulation results show that the proposed model detects whether the trend has changed in time with a small error probability and time lag. Compared with the moving-average(MA) technical indicators, the model is more accurate and less likely to identify wrong trend-changing points. We construct a trend-following strategy based on the model, and it outperforms the MA strategy in the Chinese stock market.

CO166 Room K2.41 (Hybrid 09) ADVANCES IN FACTOR MODELS AND TIME SERIES ECONOMETRICS Chair: Indeewara Perera

C0208: **Causal inference in possibly nonlinear factor models**
Presenter: **Yingjie Feng**, Tsinghua University, China

A general causal inference method is developed for treatment effects models with noisily measured confounders. The key feature is that a large set of noisy measurements are linked with the underlying latent confounders through an unknown, possibly nonlinear factor structure. The main building block is a local principal subspace approximation procedure that combines K-nearest neighbors matching and principal component analysis. Estimators of many causal parameters, including average treatment effects and counterfactual distributions, are constructed based on doubly-robust score functions. Large-sample properties of these estimators are established, which only require relatively mild conditions on the principal subspace approximation. The results are illustrated with an empirical application studying the effect of political connections on stock returns of financial firms, and a Monte Carlo experiment. The main technical and methodological results regarding the general local principal subspace approximation method may be of independent interest.

C0314: **Tackling large outliers in macroeconomic data with vector artificial neural network autoregression**
Presenter: **Yunyi Zhang**, Xiamen University, China

A regime-switching vector autoregression is developed where artificial neural networks drive time variation in the coefficients of the conditional mean of the endogenous variables and the variance-covariance matrix of the disturbances. The model is equipped with a stability constraint to ensure non-explosive dynamics. As such it is employable to account for changes in macroeconomic dynamics not only during typical business cycles but also in a wide range of extreme events, like deep recessions and strong expansions. The methodology is put to test using aggregate data for the United States that include the abnormal realizations during the recent Covid-19 pandemic. The model delivers plausible and stable structural inference and accurate out-of-sample forecasts. This performance compares favourably against a number of alternative methodologies recently proposed to deal with large outliers in macroeconomic data caused by pandemics.

C0715: **Bootstrap specification tests for dynamic conditional distribution models**
Presenter: **Mervyn Silvapulle**, Monash University, Australia
Co-authors: Indeewara Perera

Bootstrap-based tests are proposed for the specification of a given parametric conditional distribution in autoregressive time series with GARCH-type disturbances. The tests are based on an estimated residual empirical process and are implemented by parametric bootstrap. We show that the proposed tests are asymptotically valid, consistent, and have nontrivial asymptotic power against a sequence of local alternatives. Our approach relies on non-primitive regularity conditions and certain properties of exponential almost sure convergence. The regularity conditions are shown to be satisfied by GARCH(p,q); this technique of verification is applicable to other models as well. In our Monte Carlo study, the proposed tests performed well and better than several competing tests, including the information matrix test. A real data example illustrates the testing procedure.

C1025: **Mean correction and forecasting in mis-specified fractionally integrated models**
Presenter: **Kanchana Nadarajah**, University of Sheffield, United Kingdom
Co-authors: Gael Martin, Indeewara Perera, Donald Poskitt

The impact of mis-specification of short memory dynamics on estimation and forecasting in a fractionally integrated model with an unknown mean is explored. We derive the limiting distributions of three parametric estimators, namely, exact Whittle, time-domain maximum likelihood, and the conditional sum of squares (CSS), under common mis-specification of the short memory dynamics. We also show that these estimators converge to the same pseudo-true value and that their asymptotic distributions are identical to those of the frequency domain maximum likelihood and discrete Whittle (DWH) estimators. We further derived the properties of a linear predictor under mis-specification. For zero-mean processes, the linear predictor is still unbiased for the future value, but not the best predictor. In a Monte Carlo simulation study, we observe that the two time-domain estimators exhibit the smallest bias and mean squared error (MSE) as estimators of the pseudo-true value of the long memory parameter, with CSS being the most accurate estimator overall. When the mean is estimated from data, the DWH estimator is preferred in terms of bias and MSE. Furthermore, the linear predictor exhibits significant bias under mis-specification. In terms of finite sample forecast performance, CSS entails the best overall forecast error and mean squared forecast error when the mean is known, whereas the DWH exhibits the best performance when the mean is estimated.

CG037 Room Virtual R28 CONTRIBUTIONS IN FINANCIAL RISK Chair: Sandra Paterlini

C1151: **High-frequency technical analysis and systemic risk indicators**
Presenter: **Ryuichi Yamamoto**, Waseda University, Japan

A high-frequency technical analysis of individual stocks listed on the Tokyo Stock Exchange is conducted. We propose novel technical rules that derive the timing of trades according to traditional systemic risks such as shock-propagation, quote-stuffing, and tail risks measured by auto- and cross-correlations in order flows, quote-to-trade ratios, and CoVaRs. We demonstrate that both price-based technical strategies commonly used in technical analysis such as moving average rules and the newly proposed rules can exploit the significantly superior performance to the buy-and-hold rule when we trade volatile momentum or trend-reversal stocks of small-sized firms. Accordingly, this study improves stock price forecasts in high-frequency trading. The results suggest that historic prices and systemic risk indicators assist in the risk management and portfolio choices of stock investors. To the best of our knowledge, this is the first study to demonstrate the superior trading performance of individual stocks using a high-frequency technical analysis even after considering data-snooping bias and transaction costs.

C0293: **ESG measures: How they can influence the credit ratings**
Presenter: **Patrycja Chodnicka - Jaworska**, University of Warsaw, Poland

The aim is to examine the impact of the Environmental Social and Governance (ESG) measures on the credit ratings given for non-financial institutions by the two biggest credit rating agencies according to country and economic sector divisions. The hypotheses are as follows: a strong negative impact resulted from ESG risk changes on non-financial institutions credit rating changes. The reaction of the credit rating changes varied in different countries and sectors. Panel event models were used to verify the hypotheses. The study used data from the Thomson Reuters database for the period 2010-2020. The analysis was based on papers and reports on COVID-19, ESG factors and their impact on the credit rating changes, and the literature on credit rating determinants. A linear decomposition has been used for the analysis. To verify the mentioned hypotheses, long term issuer credit ratings presented by S&P, Moody and Fitch for European companies listed on the stock exchanges have been used. Credit rating changes have been collected from Thomson Reuters Database, S&P website, and companies webpages. In the analyses, financial and non-financial factors have also been used. The results suggested that during last year, the methodology presented by credit rating agencies has changed, and ESG factors are one of the basic measures that have been used to verify credit rating changes, especially those connected with the pandemic situation.

C0581: **Forecasting the climate change risk by sea-level rises using time-varying extreme value analysis**
Presenter: **Laura Garcia-Jorcano**, Universidad de Castilla-La Mancha, Spain
Co-authors: Lidia Sanchis

One of the most significant potential impacts of climate change is the sea-level rise. A better understanding and measurement of extreme sea-level rise benefits the detection and attribution of climate change signals. Using the global and regional mean sea-level rise (mm) every 10 days (Dic/1992-Oct/2020), we propose two new measures, Extreme Sea-Level Value at Rise (ExSLVaR) and Extreme Sea-Level Expected Rise (ExSLER) for forecasting extreme mean sea-level at 10 days and at 1 month calculated for 8 seas/oceans of the Earth using extreme value theory and filtered historical simulation approach. We also obtain time-varying mean sea-level rise projections for 10-80-280 years, reaching levels of 9.5 meters for 2300 for the Atlantic Ocean. The main evidence shows different regional and global forecasts. Both measures enable us to calculate the dynamic risk of extreme sea-level rise in the coastal cities/countries of these oceans/seas. Finally, we analyze the connection between our measures and financial risk in different sectors. Both measures capture more risk in sectors such as energy, and oil gas, especially in the current COVID-19 period. The results can serve as valuable inputs for these sectors in different cities/countries in deciding how much risk they are willing to accept, and consequently how much adaptation they need depending on the risk aversion of their decision-makers.

C1113: **ESGM: ESG scores and the missing pillar**
Presenter: **Ozge Sahin**, Technical University of Munich, Germany
Co-authors: Karoline Bax, Sandra Paterlini, Claudia Czado

Environmental, social, and governance (ESG) scores measure companies activities concerning sustainability and are organized on three pillars: Environmental (E-), Social (S-), and Governance (G-). Different approaches have been proposed to compute ESG scores for companies, which rely on the aggregation of many sources of information. These complementary non-financial ESG scores should provide information about the ESG performance and risks of different companies. However, the extent of missing information makes the reliability of ESG scores questionable. To account for the missing information in the underlying ESG pillars, we introduce a new pillar, the so-called Missing (M-) pillar, and propose an optimization approach to compute new ESG (ESGM) scores, which should be related to the company riskiness. The ESGM scores incorporate the extent of missing information and establish some meaningful relationship concerning the riskiness of the companies under consideration. Interesting insights into the current limitations of the ESG scoring methodology are discussed.

EO144 Room K0.50 (Hybrid 06) RECENT ADVANCES IN COMPLEX DATA ANALYSIS Chair: Juan Romo

E1321: **Neural networks interpretability through Taylor series and polynomial regression coefficients**
Presenter: **Pablo Morala**, Universidad Carlos III de Madrid, Spain
Co-authors: Rosa Lillo, Jenny Alexandra Cifuentes, Inaki Ucar

Despite being used widely, neural networks are still regarded as black boxes and are used and built through trial and error. A new approach to these problems is proposed here, by finding a relationship between a trained neural network and its weights and the coefficients of a polynomial regression that performs almost equivalently as the original neural network. To do so, Taylor expansion is used at the activation functions of each neuron. Then the resulting expressions are joint in order to obtain a combination of the original network weights that are associated with each term of a polynomial regression. The order of this polynomial regression is determined by the order used in the Taylor expansion and the number of layers in the neural network. The proposal has been empirically tested covering a wide range of different situations, showing its effectiveness and opening the door to extending this methodology. This kind of relationship between modern machine learning techniques and more traditional statistical approaches can help solve interpretability concerns. In this case, polynomial regression coefficients have a much easier interpretation than neural network weights and reduce significantly the number of parameters.

E1777: **Similarity graphs for functional data mining**
Presenter: **Raul Jimenez**, Universidad Carlos III de Madrid, Spain
Co-authors: Antonio Elias, Victor Cereijo, Juan Romo

A graphical tool is presented for functional data mining. Among several applications, the new tool can be used for clustering, classifying and outlier detection. Its implementation is easy, computationally efficient, and completely data-driven. We illustrate its functioning by performing several exploratory analyses from simulated and case-study data.

E1779: **Dimension reduction techniques based on quantiles**
Presenter: **Alvaro Mendez Civieta**, Universidad Carlos III de Madrid, Spain
Co-authors: M Carmen Aguilera-Morillo, Rosa Lillo

Partial least squares (PLS) is a well-known dimensionality reduction technique used as an alternative to ordinary least squares (OLS) in collinear or high dimensional scenarios. Being based on OLS estimators, PLS is sensitive to the presence of outliers or heavy-tailed distributions. Opposed to this, quantile regression (QR) is a technique that provides estimates of the conditional quantiles of a response variable as a function of the covariates. The usage of the quantiles makes the estimates more robust against the presence of heteroscedasticity or outliers than OLS estimators. We introduce the fast partial quantile regression algorithm (fPQR), a quantile based technique that shares the main advantages of PLS: it is a dimension reduction technique that obtains uncorrelated scores maximizing the quantile covariance between predictors and responses. But additionally, it is also a robust, quantile linked methodology suitable for dealing with outliers, heteroscedastic or heavy-tailed datasets. The median estimator of the PQR algorithm is a robust alternative to PLS, while other quantile levels can provide additional information on the tails of the responses.

E1780: **Bivariate deep kriging for large-scale spatial interpolation of wind field**
Presenter: **Pratik Nag**, King Abdullah University of Science and Technology, Saudi Arabia
Co-authors: Ying Sun, Brian Reich

High spatial resolution wind data are essential for a wide range of applications in climate, oceanographic and meteorological studies. However, these often tend to be nonGaussian with high spatial variability and heterogeneity. In spatial statistics, cokriging is commonly used for predicting bivariate spatial fields but it is not optimal except for Gaussian processes also cokriging is computationally prohibitive for large datasets. We propose a method, called bivariate DeepKriging, which is a spatially dependent deep neural network (DNN)with an embedding layer constructed by spatial Radial basis functions for bivariate spatial data prediction. We then develop a distribution-free uncertainty quantification method based on bootstrap and ensemble DNN. The proposed approach outperforms the traditional cokriging predictor with commonly used covariance functions, such as the linear model of co-regionalization and flexible bivariate Matern covariance. We show that the proposed DNN model is computationally efficient and scalable, with twenty times faster computations on average. We apply the bivariate DeepKriging method to the wind data over the Middle East region at 506771 locations. The prediction performance of the proposed method is superior over the cokriging predictors and dramatically reduces the time of computation and the large-scale computational complexity.

EO460 Room Virtual R20 STATISTICAL INFERENCE FOR CIRCULAR DATA Chair: Toshihiro Abe

E0444: **Normal-wrapped Cauchy type cylindrical distribution**
Presenter: **Toshihiro Abe**, Hosei University, Japan

A cylindrical distribution is proposed whose linear and circular parts are normal and wrapped Cauchy distributions. Its linear and circular marginal distributions, as well as conditional distributions, are also investigated. Some mathematical properties such as moment are provided. Finally, the proposed distribution is fitted to a real data set as an illustrative example.

E0445: **An extension of sine-skewed circular distributions**
Presenter: **Yoichi Miyata**, Takasaki City University of Economics, Japan

The sine skewed circular distribution is a circular probability model that can be asymmetric in shape and has the advantage that the sine and cosine moments can be written in explicit forms. We use a previous formula to propose a family of probability distributions, including the sine skewed distribution. This family includes some distributions that can give stronger asymmetry than the sine skewed distribution. Furthermore, we show that a subfamily of the distributions is identifiable with respect to parameters, and all distributions in the subfamily have explicit sine and cosine moments. We will also discuss an extension of these results to probability models on cylinders.

E0621: **Simple constructions of joint distributions with circular marginal**
Presenter: **Tomoaki Imoto**, University of Shizuoka, Japan

In diverse scientific fields, a data sample is often represented as a point in the circumference of a unit circle. Typical examples are wind direction and event time measured on a 24-hour clock. Such data are called circular data and should be modeled by a distribution defined on the circle, called circular distribution. In many cases, circular observation appears with the other circular or linear observations like wind directions at the different places, distance of animal movement and its direction, pass position and direction in soccer game and so on. For modeling and analyzing such datasets, distributions on the torus and cylinder, called toroidal distribution and cylindrical distribution, are useful tools. Several methods for constructing such distributions have been considered; maximum entropy method, trivariate reduction method, wrapping method, specified conditionals or marginals methods. Simple methods for constructing joint distributions with circular marginal are proposed as methods of specifying marginal distributions. The constructed probability density function is expressed without additional normalizing constants, and its trigonometric moments and correlation measure are expressed by simple forms, which help characterize the distribution. Other researches about estimation and application for fitting real data are also shown.

E0858: **GEL approach for robust regression on spheres**
Presenter: **Fumiya Akashi**, University of Tokyo, Japan

A nonlinear regression model is considered with a spherical predictor and a possibly heavy-tailed error term. The statistical analysis for spherical and cylindrical data is an important topic in the fields of seismic wave analysis, analysis for orientation of wildfire, wind direction analysis, and so on. We make use of a least absolute deviation-based generalized empirical likelihood (GEL) approach to construct a robust statistic for heavy-tailed observations. The proposed GEL statistics is shown to have a pivotal limit distribution, and based on the asymptotic result, we also propose a method for the construction of confidence intervals for the nonlinear regression function at a certain direction. The finite sample performance of the proposed method is investigated by some simulation experiments.

E1746: **Pair circulas modelling for circular multivariate time series**
Presenter: **Hiroaki Ogata**, Tokyo Metropolitan University, Japan

Modelling multivariate circular time series is considered. The cross-sectional and serial dependence is described by circulas, which are analogs of copulas for circular distributions. Due to a simple expression of the dependence structure, we decompose a multivariate circula density to a product of several pair circula densities. Moreover, in order to reduce the number of pair circula densities, we consider strictly stationary multi-order Markov processes. Some simulation studies are provided to see the behavior of the proposed model.

EO475 Room Virtual R21 STATISTICS FOR HILBERT SPACES Chair: Gil Gonzalez-Rodriguez

E0912: **Flexible beta regression with functional covariates: A Bayesian approach**
Presenter: **Agnese Maria Di Brisco**, Universita del Piemonte Orientale, Italy
Co-authors: Enea Bongiorno, Aldo Goia, Sonia Migliorati

Modeling bounded continuous response variables, such as rates and proportions, is a common issue in many disciplines. Due to the constraint on the response, possible solutions are the beta and the flexible beta regression models. The latter has been recently proposed and it is based on a special mixture of betas designed to cope with (not limited to) bimodality, heavy tails, and outlying observations. These models are generalized to the case of Hilbert valued covariates. Estimation issues are dealt with through a combination of standard basis expansion and MCMC techniques. Specifically, we propose to select the most significant coefficients of the expansion through Bayesian variable selection methods that take advantage of shrinkage priors. The effectiveness of the proposal is illustrated by using numerical examples and an application of spectrometric analysis on milk specimens.

E0966: **Depth-based two-sample testing**
Presenter: **Felix Gnettner**, Otto-von-Guericke-Universitaet Magdeburg, Germany
Co-authors: Claudia Kirch, Alicia Nieto-Reyes

Depth functions provide measures of the deepness of a point with respect to a given set of observations. This non-parametric concept can be applied in spaces of any dimension and entails a center-outward ordering for the given data. A two-sample test has been previously proposed that is based on depth-ranks and offers opportunities for further investigations: Observing that the corresponding test statistic $Q(X,Y)$ is not symmetric with respect to the two samples X and Y, the power can be greatly increased if $Q(X,Y)$ and $Q(Y,X)$ are jointly considered. Within the last years, depths with respect to functional data have been established that we combine with this procedure to obtain new non-parametric two-sample tests for functional data. We investigate the asymptotic behaviour of this modified test procedure for several classes of depths including depths for functional data.

E0998: **Generalized functional additive mixed models with compositional covariates for Spanish Covid-19 incidence data**
Presenter: **Sonja Greven**, Humboldt University of Berlin, Germany
Co-authors: Matthias Eckardt, Jorge Mateu

A generalized functional additive mixed model is extended to the situation when the outcomes are functions and parts of the independent variables are finite or infinite compositions, i.e. functional compositions, carrying relative information of a whole. Relying on the isometric isomorphism of the Bayes Hilbert space of probability densities and the space of square-integrable functions with integration-to-zero constraint through the centred log-ratio (clr) transformation, functional compositions are incorporated as functional covariates into the model using a flexible basis function representation that also accounts for the integration-to-zero constraint. The extended generalized functional additive mixed model allows for the estimation of linear, nonlinear and also time-varying effects of scalar and functional covariates, as well as of the effects of (potentially spatial) grouping factors, in addition to the compositional effects. The potential of the extended model is shown by estimating the effect of age curves, i.e. functional compositions, and smoking status on regional Covid-19 incidence data for Spain.

E1196: **Nonparametric regression and classification with functional, categorical, and mixed covariates**
Presenter: **Jan Gertheiss**, Helmut Schmidt University, Germany
Co-authors: Leonie Selk

Nonparametric prediction with multiple covariates is considered, in particular categorical or functional predictors (from a Hilbert space, such as the space of square-integrable functions), or a mixture of both. A linear combination of distance measures each calculated on single covariates is proposed, with weights being estimated from the training data. Emphasis is put on the case of a categorical, multi-class response, because the number of corresponding, nonparametric methods found in the literature that can be used with multiple categorical/functional predictors is very limited. The methodology resented is illustrated and evaluated on both artificial and real world data. Particularly it is observed that prediction accuracy can be increased, and irrelevant, noise variables can be identified/removed by "downgrading" the corresponding distance measures.

E1645: **Simultaneous inference for function-valued parameters**
Presenter: **Dominik Liebl**, University Bonn, Germany
Co-authors: Matthew Reimherr

A new approach for constructing simultaneous confidence bands for function-valued parameters is presented. The bands are fast to compute as they are based on nearly closed-form expressions and, therefore, do not require computationally expensive resampling based methods. The shape of the bands can be constructed according to a desired criterion specified by the user. A particularly interesting criteria is the proposed concept of "fair" or equitable bands, which leads to simultaneous confidence bands that have an adaptive width reflecting the local multiple testing problem. Our bands are constructed by integrating and extending tools from Random Field Theory, an area that has yet to overlap with Functional Data Analysis. Large sample properties are analyzed using asymptotic statistics, and finite-sample properties of our method are investigated by means of extensive simulation studies. Applicability is demonstrated using different real data applications from sports biomechanics and economics.

EO509 Room Virtual R22 MODEL SPECIFICATION TESTS Chair: Bojana Milosevic

E0556: **Goodness-of-fit testing for normal mixtures**
Presenter: **Dimitrios Bagkavos**, University of Ioannina, Greece

A novel goodness-of-fit test is introduced for the case where the hypothesized distribution is a mixture of normal distributions. The theoretical results contributed include analytic quantification of the test statistic asymptotic distribution under both the null and the alternative hypothesis and closed-form expressions for its power under Pitman alternatives, assuming a parametrically estimated underlying model. Further, the Edgeworth

expansion of the proposed test's size and power functions are derived and employed in developing a bandwidth selector, designed to optimize power, subject to keeping the size constant. Numerical examples illustrate the performance of the proposed test in practice.

E0665: **Specification tests for selection bias models under random truncation**
Presenter: **Jacobo de Una-Alvarez**, University of Vigo, Spain

Random truncation means that the target random variable is observed only when it falls within a random set. This generally results in a sampling bias, in the sense that different values of the target may have different chances to be sampled. We will consider the problem of testing for the null hypothesis of ignorable sampling bias. When the null is true, ordinary estimators are consistent and, therefore, no correction for random truncation is needed. Two different test statistics based on the NPMLE of the sampling probability will be introduced, and their distributional convergence under the null will be justified. Bootstrap algorithms to approximate the null distribution of the test will be presented. Applications to specific forms of truncation will be given. Simulation results and real data analyses will be provided. Possible extensions to general selection bias models will be discussed.

E0695: **Testing normality of a large number of populations**
Presenter: **Maria Dolores Jimenez-Gamero**, Universidad de Sevilla, Spain

Suppose we have independent samples coming from k populations. We consider the problem of (simultaneously) testing that the populations are normal. The means and the variances may be different. The number of populations is allowed to increase to infinite. It may be even larger than the sample sizes. In our developments, it is assumed that the data is univariate in order to keep the notation simple. Nevertheless, the results also apply (with minor modifications) to multivariate data.

E0947: **On a new omnibus test of fit based on a characterisation of the uniform distribution**
Presenter: **Jaco Visagie**, North-West University, South Africa
Co-authors: Bruno Ebner, Shawn Liebenberg

The classical goodness-of-fit problems for univariate distributions is revisited. We propose a new testing procedure based on a characterisation of the uniform distribution. Asymptotic theory for the simple hypothesis case is provided in a Hilbert-Space setting, including the asymptotic null distribution as well as values for the first four cumulants of this distribution, which are used to fit a Pearson system of distributions as an approximation to the limit distribution. Numerical results indicate that the null distribution of the test converges quickly to its asymptotic distribution, making the critical values obtained using the Pearson system particularly useful. Consistency of the test is shown against any fixed alternative distribution and we derive the limiting behaviour under fixed alternatives with an application to power approximation. We demonstrate the use of the newly proposed test when testing composite hypotheses. A Monte Carlo power study compares the finite sample power performance of the newly proposed test to existing omnibus tests in both the simple and composite hypothesis settings. This power study includes results related to testing for the uniform, normal and Pareto distributions. The empirical results obtained indicate that the test is competitive. An application of the newly proposed test in financial modelling is also included.

E0949: **Logistic or not logistic?**
Presenter: **James Allison**, Northwest University, South Africa
Co-authors: Bruno Ebner, Marius Smuts

A new class of goodness-of-fit tests is proposed for the logistic distribution based on a characterisation related to the density approach in the context of Stein's method. This characterisation based test is a first of its kind for the logistic distribution. The asymptotic null distribution of the test statistic is derived and it is shown that the test is consistent against fixed alternatives. The finite sample power performance of the newly proposed class of tests is compared to various existing tests by means of a Monte Carlo study. It is found that this new class of tests are especially powerful when the alternative distributions are heavy-tailed, like Student's t and Cauchy, or for skew alternatives such as the log-normal, gamma and chi-square distributions.

EO639 Room Virtual R23 INDEPENDENCE TESTS, VARIABLE SELECTION, AND ROBUST CLASSIFICATION Chair: Yuexiao Dong

E0214: **Fast distance correlation chi-square test**
Presenter: **Cencheng Shen**, University of Delaware, United States

Distance correlation has gained much recent attention in the data science community: the sample statistic is straightforward to compute and asymptotically equals zero if and only if there is independence, making it an ideal choice to discover any type of dependency structure given sufficient sample size. One major bottleneck is the testing process: because the null distribution of distance correlation depends on the underlying random variables and metric choice, it typically requires a permutation test to estimate the null and compute the p-value, which is very costly for a large amount of data. To overcome the difficulty, we propose a chi-square test for distance correlation. Method-wise, the chi-square test is non-parametric, extremely fast, and applicable to bias-corrected distance correlation using any strong negative type metric or characteristic kernel. The test exhibits a similar testing power as the standard permutation test, and can be utilized for K-sample and partial testing. Theory-wise, we show that the underlying chi-square distribution well approximates and dominates the limiting null distribution in the upper tail, prove the chi-square test can be valid and universally consistent for testing independence, and establish a testing power inequality with respect to the permutation test.

E1142: **Testing the linear mean and constant variance conditions in sufficient dimension reduction**
Presenter: **Yuexiao Dong**, Temple University, United States

Sufficient dimension reduction methods characterize the relationship between the response Y and the covariates X, through a few linear combinations of the covariates. Extensive techniques are developed, among which the inverse regression-based methods are perhaps the most appealing in practice because they do not involve multi-dimensional smoothing and are easy to implement. However, these inverse regression-based methods require two distributional assumptions on the covariates. In particular, the first-order methods, such as the sliced inverse regression, require the linear conditional mean (LCM) assumption, while the second-order methods, such as the sliced average variance estimation, require additionally the constant conditional variance (CCV) assumption. We propose to check the validity of the LCM and the CCV conditions through mean independence tests, which are facilitated by the martingale difference divergence. We suggest a consistent bootstrap procedure to decide the critical values of the test. Monte Carlo simulations as well as an application to the horse mussels dataset are conducted to demonstrate the finite sample performance of our proposal.

E1274: **On sufficient variable screening using log odds ratio filter**
Presenter: **Wenbo Wu**, University of Texas at San Antonio, United States

A new dependence measure is proposed which is called the log odds ratio statistic to be used under the sufficient variable screening framework. In addition, we propose an ensemble variable 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. We establish the sure screening properties of the fused log odds ratio filter for both marginal variable screening and sufficient variable screening.

E1392: **Multivariate functional group sparse regression: Functional predictor selection**
Presenter: **Jun Song**, Korea University, Korea, South

A method is presented for functional predictor selection and the estimation of smooth functional coefficients simultaneously in a scalar-on-function regression problem under a high-dimensional multivariate functional data setting. In particular, we develop a method for functional group-sparse

regression under a generic Hilbert space of infinite dimension. Then we show the convergence of algorithms and the consistency of the estimation and selection under infinite-dimensional Hilbert spaces. Simulation and fMRI data application will be presented at the end to show the effectiveness of the methods in both the selection and estimation of functional coefficients.

E1381: **Robust envelope discriminant analysis**
Presenter: **Abdul-Nasah Soale**, University of Notre Dame, United States

Classical linear discriminant analysis imposes an assumption of normality on the conditional distribution of the predictors given the classes. However, in practice, this assumption is easily violated. Motivated by the recent work of envelope discriminant analysis, our robust envelope proposal extends the estimation of the envelope discriminant subspace beyond the normality assumption. We demonstrate the promising performance of our proposal using both synthetic and real data analysis.

EO687 Room Virtual R24 ADVANCES IN VARIATIONAL APPROXIMATIONS Chair: Luca Maestrini

E0839: **Sparse linear mixed model selection via streamlined variational Bayes**
Presenter: **Emanuele Degani**, University of Padua, Italy
Co-authors: Luca Maestrini, Dorota Toczydlowska, Matt P Wand

Variational approximations facilitate fast approximate inference for the parameters of a variety of statistical models. However, for mixed models having a high number of random effects, simple application of standard variational inference principles does not lead to fast approximate inference algorithms, due to the size of model design matrices and inefficient treatment of sparse matrix problems arising from the required approximating density parameters updates. We illustrate how previous streamlined variational inference procedures can be generalized to make fast and accurate inferences for the parameters of linear mixed models with nested random effects and priors for selecting fixed effects. The variational inference algorithms achieve convergence to the same optima of their standard implementations, but with significantly lower computational effort, memory usage, and time, especially for large numbers of random effects. Using simulated and real data examples, we assess the quality of automated procedures for fixed effects selection that only rely upon variational approximations and are free from hyperparameters tuning, and also show high accuracy of the approximations against Markov Chain Monte Carlo.

E0776: **Bayesian functional principal components analysis via variational message passing**
Presenter: **Tui Nolan**, University of Cambridge, United Kingdom
Co-authors: Jeff Goldsmith, David Ruppert

Standard approaches for functional principal components analysis rely on an eigendecomposition of a smoothed covariance surface in order to extract the orthonormal eigenfunctions representing the major modes of variation in a set of functional data. This approach can be a computationally intensive procedure, especially in the presence of large datasets with irregular observations. We outline a variational Bayesian approach, which aims to determine the Karhunen-Loeve decomposition directly without smoothing and estimating a covariance surface. More specifically, we incorporate the notion of variational message passing over a factor graph because it removes the need for rederiving approximate posterior density functions if there is a change in the model. Instead, model changes are handled by changing specific computational units, known as fragments, within the factor graph. Indeed, this is the first method to address a functional data model via variational message passing. Our approach introduces two new fragments that are necessary for Bayesian functional principal components analysis. We present the computational details, a set of simulations for assessing the accuracy and speed of the variational message passing algorithm and an application to the United States temperature data.

E0362: **Use of model reparametrization to improve variational Bayes**
Presenter: **Siew Li Linda Tan**, National University of Singapore, Singapore

Using model reparametrization is proposed to improve variational Bayes inference for hierarchical models whose variables can be classified as global (shared across observations) or local (observation specific). Posterior dependence between local and global variables is minimized by applying an invertible affine transformation on the local variables. The functional form of this transformation is deduced by approximating the posterior distribution of each local variable conditional on the global variables by a Gaussian density via a second-order Taylor expansion. Variational Bayes inference for the reparametrized model is then obtained using stochastic approximation. The approach can be readily extended to large datasets via a divide and recombine strategy. Using generalized linear mixed models, we demonstrate that reparametrized variational Bayes (RVB) provides improvements in both accuracy and convergence rate compared to state of the art Gaussian variational approximation methods.

E0751: **A new variational family for Bayesian deep learning**
Presenter: **Susan Wei**, University of Melbourne, Australia

Unlike in regular statistical models, the posterior distribution over neural network weights is not asymptotically Gaussian. As established in singular learning theory, the posterior distribution over the parameters of a singular model is, asymptotically, a mixture of standard forms. Loosely, this means the parameter space can be partitioned such that in each local parameter set, the average log-likelihood ratio can be made normal crossing via an algebraic-geometrical transform known as a resolution map. We leverage this under-appreciated result to propose a new variational family for Bayesian deep learning. Affine coupling layers are employed to learn the unknown resolution map, effectively rendering the proposed methodology a normalizing flow with the generalized gamma as the source distribution, rather than the multivariate Gaussian typically employed.

E0577: **Non-stationary Gaussian process discriminant analysis with variable selection for high-dimensional functional data**
Presenter: **Weichang Yu**, University of Melbourne, Australia
Co-authors: Sara Wade, Howard Bondell, Lamiae Azizi

High-dimensional classification and feature selection problems are ubiquitous with the recent advancement in data acquisition technology. In several application areas such as biology, genomics and proteomics, the analysed data are often functional and have a complex structure. The high dimensionality of the data coupled with the correlation structure poses serious challenges to the data analysis. Many existing statistical and machine learning models either fit the data poorly or suffer from a lack of model interpretability. We propose a novel Bayesian discriminant analysis-based model that addresses these challenges in a unified framework and performs variable selection simultaneously. The model is a two-layer non-stationary Gaussian process to model the complex high-dimensional observations coupled with an Ising prior to identify differentially-distributed locations. The model inference scalability is achieved via developing a variational scheme that exploits advances in the use of sparse structures covariance matrices. We show the performance of our proposed model in simulated datasets and various proteomics-related mass spectrometry real datasets (breast cancer and SARS-CoV-2). Moreover, we demonstrate how the output from our proposed model may be used to address scientific hypotheses, offering explainability as well as uncertainty quantification, which are crucial to increase trust and social acceptance of data-driven tools.

EO820 Room Virtual R25 HIGH-DIMENSIONAL REGRESSION MODELS Chair: Zhaoyuan Li

E0234: **Extension of the Lagrange multiplier test for error cross-section independence to large panels with non-normal errors**
Presenter: **Zhaoyuan Li**, The Chinese University of Hong Kong, Shenzhen, China
Co-authors: Jeff Yao

The aim is to reexamine the seminal Lagrange multiplier test for cross-section independence in a large panel model where both the number of cross-sectional units n and the number of time-series observations T can be large. The first contribution is an enlargement of the test with two

extensions: firstly, the new asymptotic normality is derived in a simultaneous limiting scheme where the two dimensions (n,T) tend to infinity with comparable magnitudes; second, the result is valid for general error distribution (not necessarily normal). The second contribution is a new test statistic based on the sum of the fourth powers of cross-section correlations from OLS residuals, instead of their squares used in the Lagrange multiplier statistic. This new test is generally more powerful, and the improvement is particularly visible against alternatives with weak or sparse cross-section dependence. Both simulation study and real data analysis are proposed to demonstrate the advantages of the enlarged Lagrange multiplier test and the power enhanced test compared to the existing procedures.

E0487: **KOO: A scalable model selection rules in high-dimensional regression**
Presenter: **Jiang Hu**, Northeast Normal University, China
Co-authors: Zhidong Bai, Kwok Pui Choi, Yasunori Fujikoshi

Variable selection is essential for improving inference and interpretation in multivariate linear regression. We consider the strong consistency of the high-dimensional knock-one-out (KOO) methods. The proposed method removes the penalties while simultaneously reducing the conditions for the dimensions and sizes of the regressors. Simulation studies and real data analysis support our conclusions.

E0574: **Selective confidence intervals for martingale regression model**
Presenter: **Ka Wai Tsang**, The Chinese University of Hong Kong, Shenzhen, China

The problem of constructing confidence intervals is considered for selected coefficients in a martingale regression model. It is a common practice in statistical analysis to perform data-driven variable selection and derive statistical inference from the resulting model. Such inference enjoys none of the guarantees that classical statistical theory provides for confidence intervals when the model has been chosen a priori. To overcome the inherent difficulties of post-selection confidence intervals, we introduce consistent estimators of the selected parameters and work on a resampling approach.

E0900: **On estimating generalization gaps via the functional variance in overparameterized models**
Presenter: **Keisuke Yano**, The Institute of Statistical Mathematics, Japan
Co-authors: Akifumi Okuno

The focus is on a generalization gap estimation for overparameterized models such as deep neural networks. We show that the functional variance, a key concept in defining a widely-applicable information criterion, well approximates the difference between the generalization error and an empirical error in overparameterized linear regression models. Overparametrerized linear regression models arise by considering regimes where deep neural networks can be well-approximated by linear functions of their parameters; for example, the neural tangent kernel regime. For practical implementation, we propose a Langevin approximation of the functional variance, which can be implemented consistently with gradient-based optimization algorithms and leverages only the first-order gradient of a loss function. Through numerical experiments, we demonstrate the applicability of the Langevin functional variance for overparameterized models.

E0460: **A novel computationally scalable high-dimensional vector autoregressive moving average model**
Presenter: **Yao Zheng**, University of Connecticut, United States
Co-authors: Feiqing Huang, Guodong Li, Kexin Lu

Classical VARMA models are very popular in modeling general linear processes due to their parsimony and favorable forecasting performance. Yet, the complicated identification issue and heavy computational burden hinder their practicality in the high dimensional regime. We introduce a scalable autoregressive moving average (SARMA) model that inherits the VARMA model's interpretability and the rich, dynamic structure of the VARMA model while avoiding the identification problem. Most notably, this family of multivariate linear processes contains virtually all VARMA processes and can also be easily extended to cover other forms of linear processes. In the high dimensional regime, we propose a low-Tucker-rank approach for further dimension reduction. Non-asymptotic error bounds for this model are derived, and a computationally scalable algorithm is developed. Simulation and real data analysis demonstrate the advantages of the proposed SARMA approach over existing methods.

EO260 Room Virtual R26 INFERENCE FOR NON-REGULAR STOCHASTIC PROCESSES Chair: Kengo Kamatani

E0479: **Gaussian quasi-information criterion for ergodic SDEs**
Presenter: **Shoichi Eguchi**, Osaka Institute of Technology, Japan
Co-authors: Hiroki Masuda

There are several studies of model selection for stochastic differential equations (SDEs), for example, the contrast-based information criterion for ergodic diffusion processes and the Schwarz type information criterion for locally asymptotically quadratic models. We consider pure-jump Lévy noise-driven SDEs as the candidate models and propose the AIC-type information criterion the stepwise model selection procedure.

E0864: **Local asymptotic normality for ergodic jump diffusion processes**
Presenter: **Yuma Uehara**, Kansai University, Japan
Co-authors: Teppei Ogihara

Sufficient conditions for local asymptotic mixed normality are studied in order to deal with a wider class of statistical models. Moreover, we show that the local asymptotic mixed normality of a statistical model generated by approximated transition density functions is implied for the original model. Together with a density approximation by means of thresholding techniques, we derive the local asymptotic normality for a statistical model of discretely observed jump-diffusion processes where the drift coefficient, diffusion coefficient, and jump structure are parametrized. As a consequence, the quasi maximum likelihood and Bayes type estimators are shown to be asymptotically efficient in this model.

E0971: **Local asymptotic normality property for fractional Brownian motion with measurement errors**
Presenter: **Tetsuya Takabatake**, Hiroshima University, Japan

The aim is to prove a Local Asymptotic Normality (LAN) property for a fractional Brownian motion (fBm) observed at equidistant time-points in the presence of measurement errors given by some Gaussian moving average processes. We will show that the LAN property for an fBm with measurement errors heavily depends on (1) a limit of the ratio between the variances of the increments of fBm and measurement errors when the sample size goes to infinity, (2) an asymptotic behavior of the ratio between spectral density functions of the increments of fBm and measurement errors at the origin.

E1109: **An M-estimatior for stochastic differential equations driven by fractional Brownian motion with small Hurst parameter**
Presenter: **Kohei Chiba**, Osaka University, Japan

A stochastic differential equation driven by a fractional Brownian motion with small Hurst parameter is considered. We are interested in estimating the drift parameter from the completely observed data. We propose an M-estimator for the drift parameter. Under some assumptions on the drift coefficient, our estimator has consistency, asymptotic normality and moment convergence property.

E0443: **Adaptive maximum likelihood type estimators for discretely observed small diffusion processes**
Presenter: **Masayuki Uchida**, Osaka University, Japan

Parameter estimation is considered for multi-dimensional diffusion processes with small dispersion parameters from discrete observations. The joint estimation of both the drift and diffusion parameters of diffusion processes with small dispersion parameters was previously investigated under the general conditions on the sample size and the small dispersion parameter. We propose two kinds of adaptive maximum likelihood type

estimators for both the drift and diffusion parameters of diffusion processes with small dispersion parameters from the viewpoint of numerical computation. It is also shown that the proposed estimators of both the drift and diffusion parameters have asymptotic normality under the same general conditions on the sample size and the small dispersion parameter as previously. Furthermore, some examples and simulation results of the proposed drift and diffusion parameters estimators are given.

EO716 Room Virtual R27 EXTREMES AND APPLICATIONS Chair: Boris Beranger

E0658: **High-dimensional modeling of spatial and spatio-temporal conditional extremes using INLA and the SPDE approach**
Presenter: **Thomas Opitz**, BioSP-INRAE, France
Co-authors: Emma Simpson, Jenny Wadsworth

The conditional extremes framework allows for event-based stochastic modeling of dependent extremes, and has recently been extended to spatial and spatiotemporal settings. After standardizing the marginal distributions and applying an appropriate linear normalization, certain nonstationary Gaussian processes can be used as asymptotically motivated models for the process conditioned on threshold exceedances at a fixed reference location. We adopt a Bayesian perspective by implementing estimation through the integrated nested Laplace approximation (INLA), allowing for novel and flexible semi-parametric specifications of the Gaussian mean function. By using Gauss-Markov approximations of the Matern covariance function (known as the Stochastic Partial Differential Equation approach) at a latent stage of the model, likelihood-based inference becomes feasible even with thousands of observed locations. We explain how constraints on the spatial and spatiotemporal Gaussian processes, arising from the conditioning mechanism, can be implemented through the latent variable approach without losing the computationally convenient Markov property. We discuss model comparison and illustrate the approach with gridded Red Sea surface temperature data at over 6,000 observed locations. Posterior sampling is exploited to study the probability distribution of cluster functionals of spatial and spatiotemporal extreme episodes.

E0746: **Road safety of passing maneuvers: A bivariate extreme value theory approach under non-stationary conditions**
Presenter: **Ana Ferreira**, IST-ID, Portugal

Observed accidents have been the main resource for road safety analysis over the past decades. Although such reliance seems quite straightforward, the rare nature of these events has made safety difficult to assess, especially for new and innovative traffic treatments. Surrogate measures of safety have allowed to step away from traditional safety performance functions and analyze safety performance without relying on accident records. In recent years, the use of extreme value theory (EV) models in combination with surrogate measures to estimate accident probabilities has gained popularity within the safety community. We extend existing efforts on EV for accident probability estimation using two dependent surrogate measures. Using detailed trajectory data from a driving simulator, we model the joint probability of head-on and rear-end collisions in passing maneuvers. In the estimation, we account for driver-specific characteristics and road infrastructure variables. We show that accounting for these factors improves the head-on collision probability estimation. We also present an exploratory structure and results for combining surrogate measures that describe correlated events: in our case of passing maneuvers this considers the joint distribution of head-on and rear-end collision. Such a feature is essential to keep up with the expectations from surrogate safety measures for the integrated analysis of accident phenomena.

E0874: **Empirical Bayes analysis of maxima**
Presenter: **Stefano Rizzelli**, Catholic University - Milan, Italy
Co-authors: Simone Padoan

Predicting future observations is the central goal of several statistical applications concerning extreme value data. Under mild assumptions, extreme-value theory justifies modelling linearly normalized sample maxima by max-stable distributions. The Bayesian paradigm offers a direct approach to forecasting and uncertainty quantification. Various proposals for Bayesian inferential procedures have been formulated in recent years, though they typically disregard the asymptotic bias inherent in the use of max-stable models, incorporating no information on norming sequences in prior specifications for scale and location parameters. We propose an empirical Bayes approach that suitably addresses this point via data-dependent priors. We illustrate the resulting asymptotic posterior concentration properties and pinpoint their implications for the estimation and prediction of future observations.

E0383: **A sparse Gaussian scale mixture process for short-range tail dependence and long-range independence**
Presenter: **Raphael Huser**, King Abdullah University of Science and Technology, Saudi Arabia
Co-authors: Arnab Hazra

Various natural phenomena, such as precipitation, exhibit short-range spatial tail dependence. However, the available models in the spatial extremes literature generally assume that spatial tail dependence persists across the entire spatial domain. We develop a novel Bayesian Gaussian scale mixture model, where the Gaussian process component is driven by a stochastic partial differential equation that yields a sparse precision matrix, and the random scale component is modeled as a low-rank Pareto-tailed or Weibull-tailed spatial process determined by compactly supported basis functions. We show that our model is tail-stationary, and we demonstrate that it can capture a wide range of tail dependence structures as a function of distance, such as strong tail dependence at short distances and tail independence at large distances. The sparse structure of our spatial model allows fast Bayesian computation, even in high spatial dimensions. Our inference approach relies on a well-designed Markov chain Monte Carlo algorithm. In our application, we fit our model to analyze heavy monsoon rainfall data in Bangladesh. Numerical experiments show that our model provides a good fit to the data. It can be exploited to draw inferences on long-term return levels for marginal rainfall at each site, and for spatial aggregates.

E1763: **Narrowing the gap between theory and practice for modelling rainfall extremes**
Presenter: **Kate Saunders**, QUT, Australia

The challenge for modelling rainfall extremes using many extreme value methods is that there exists a gap between the theory and how reliably the theory is representing the true physical process. Of particular concern here is the common practice of separating the bulk of the data from the exceedances to fit models of extreme rainfall. In a univariate context, this might be acceptable, but as extreme value theory methods evolve to support spatial-temporal approaches to modelling rainfall extremes, continuing to ignore the bulk of the data may result in unintended consequences. The most concerning is that we may fail to reliably estimate the risk posed by extreme rainfall events. We consider inter-disciplinary perspectives on modelling rainfall extremes and explore where the methods in extreme value theory need to evolve to meet the needs of the application. We consider domain knowledge from the compound event community, data science community and climate community, and discuss how this domain knowledge can be used to build extreme value models that more reliably represent the risk posed by extreme rainfall events.

EO440 Room Virtual R30 RECENT DEVELOPMENTS IN MULTIVARIATE DATA ANALYSIS Chair: Anne Ruiz-Gazen

E1045: **Impact of linear constraints on a multivariate binary classification problem**
Presenter: **Sonia Perez-Fernandez**, University of Oviedo, Spain

The classification accuracy of a continuous variable - frequently called a marker - to distinguish between two groups is usually measured in terms of the sensitivity and the specificity, which are the probabilities of correctly classifying a subject from each group. When classification rules are based on a unique threshold for the marker, those resulting probabilities for every possible cut-off point are frequently displayed on a single graphic, called the Receiver Operating Characteristic (ROC) curve. There are some generalizations of the ROC curve to accommodate scenarios where using a unique cut-off point is far from the optimal classification rule. One example is the so-called general ROC (gROC) curve, where two thresholds are considered. A reformulation of the definition of the ROC curve is combined with the idea underlying the gROC curve, with the

goal of constructing interpretable classification regions which report the maximum sensitivity for a particular specificity. The study is carried out in a conditionally normal bivariate scenario. The resulting decision rules and ROC curves are compared to the optimal ones without imposing any restrictions.

E1246: **Complex-valued covariance models for spatio-temporal vector data**
Presenter: **Sandra De Iaco**, University of Salento, Italy

In geostatistical literature, there are various contributions focused only on modeling the spatial evolution of vector data with two components in the framework of the theory of complex-valued random fields. However, the temporal perspective is analyzed separately or used to model time-varying complex covariance models. Thus, in this context, it is surely challenging to propose some advances in modeling the joint spatial and temporal behavior of vector data with a reasonable representation on a complex domain. After introducing the theoretical background regarding the complex formalism of a spatio-temporal random field, some techniques for building new families of spatio-temporal models are discussed. Then, the spatio-temporal complex modeling is applied to sea current data referred to the US East and Gulf Coast. The results regarding a comparative analysis between different complex-valued covariance models are also presented.

E1255: **Blind source separation for non-stationary random fields**
Presenter: **Klaus Nordhausen**, University of Jyvaskyla, Finland
Co-authors: Christoph Muehlmann, Francois Bachoc

Multivariate spatial data possess many challenges in proper statistical modelling. Firstly, dependencies need to be modelled not only on-sight but as a function of spatial separation. Secondly, as the data is multivariate cross-dependencies need to be considered as well. Usually, this is done in terms of a cross-covariance function, where the multivariate random field is supposed to fulfil second-order stationarity assumptions. Spatial Blind Source Separation (SBSS) is a recently introduced unsupervised statistical tool, which is designed to deal with the challenges of multivariate covariance modelling. In the SBSS model, it is assumed that the observable random field is formed by a latent variable linear mixture, where the latent random field is formed by uncorrelated, weakly stationary random fields. However, second-order stationarity might be too restricting when for example the spatial domain increases and the on-sight variances vary in space, or the spatial covariance actually depends on the locations themselves rather than the distances between them. This leads to non-stationary spatial modeling which again possesses the same challenges as described above and additionally opens the door for a variety of second-order stationary violations. We address these issues by combining the principles of Blind Source Separation and non-stationary geostatistics which leads to Spatial Non-stationary Source Separation (SNSS).

E1366: **New algorithms for implementing invariant component selection**
Presenter: **Aurore Archimbaud**, Toulouse School of Economics, France
Co-authors: Zlatko Drmac, Klaus Nordhausen, Una Radojicic, Anne Ruiz-Gazen

ICS (Invariant Component Selection) is a method for multivariate analysis based on the joint diagonalization of two scatter matrices. In the statistics literature, it is usually just stated that it can be solved as a generalized eigenvalue-eigenvector problem and little details about the actual computations are given. Some implementations in R are for example in the R packages ICS and ICtest. However, the existing functions do not take into account singular or nearly singular scatter matrices. In particular, in the case of nearly singular matrices, the existing algorithms do work but may encounter numerical instability. The aim is to propose alternative implementations of ICS and compare them to the existing ones.

EO190 Room Virtual R36 DIRECTIONAL STATISTICS IN MULTIDISCIPLINARY DOMAINS Chair: Andriette Bekker

E0183: **Directional statistics and protein bioinformatics: Recent developments**
Presenter: **Christophe Ley**, Ghent University, Belgium

In the bioinformatics field, there has been a growing interest in modelling dihedral angles of amino acids by viewing them as data on the torus. This has motivated, over the past years, new proposals of distributions on the bivariate torus. More generally, questions from the protein structure prediction problem have generated quite some research activity in the field of directional statistics. An overview of the generated research questions and novel statistical methods that have arisen thanks to the demands from protein bioinformatics will be presented.

E0345: **Inference for toroidal models with circula densities generated by Fourier series**
Presenter: **Arthur Pewsey**, University of Extremadura, Spain

Inference is considered for bivariate circular models derived using a marginal specification construction in combination with highly tractable circula densities generated through different patterns of non-zero Fourier coefficients. Specifically, a highly successful model identification tool and methods for parameter estimation and goodness-of-fit testing are introduced. The results from a numerical experiment comparing the modelling capabilities of such bivariate circula densities and five existing models are presented. The application of the different models is illustrated in an analysis of wind directions. An important advantage of our approach is that it facilitates the separate modelling of the circular marginals and a circula density in a structured sequential way. It also accommodates formal goodness-of-fit testing, an issue neglected in the literature related to the application of existing models for toroidal data.

E0672: **Mixtures of Kato-Jones distributions on the circle with an application to traffic count data**
Presenter: **Shogo Kato**, Institute of Statistical Mathematics, Japan
Co-authors: Kota Nagasaki, Wataru Nakanishi

Kato-Jones distribution is a probability distribution on the circle that is unimodal and affords a wide range of skewness and kurtosis. Motivated by a multimodal skewed data set which appears in traffic engineering, mixtures of Kato-Jones distributions are considered. A key reparametrization is done to achieve the identifiability of the proposed mixtures. With this reparametrization, two methods for parameter estimation, namely, a modified method of moments and the maximum likelihood method, are presented. The modified method of moments estimation is relatively fast and provides the reasonable initial value of the algorithm for the maximum likelihood estimation. The maximum likelihood estimation can be done using the EM algorithm. These two methods are seen to be useful for fitting the proposed mixtures to the traffic counter data set of interest.

E0926: **A unifying generator for cardioid modelling in circular statistics**
Presenter: **JT Ferreira**, University of Pretoria, South Africa
Co-authors: Mohammad Arashi, Andriette Bekker, Najmeh Nakhaeirad, Delene van Wyk

The cardioid distribution maintains a fundamental position within circular statistics, and has been well studied and applied for analyzing circular data and in regression contexts. Theoretical properties of the cardioid have been derived and implemented to offer useful applications. An alternative construction to derive a family of circular distributions is reported which incorporates the kernel of the cardioid as a generator, and special cases contained in this family are highlighted. Particular emphasis is on a specific member, termed cardioid-t, and synthetic data investigations, as well as real data, illustrates the candidacy of this model in the directional statistics environment.

E1005: **Establishing a Bayesian nonparametric density estimation for biased circular data**
Presenter: **Najmeh Nakhaeirad**, University of Pretoria, South Africa
Co-authors: Andriette Bekker, Mohammad Arashi

Circular data can be recorded with some errors in variables. This results in biased data. Routine directional methods fail to correctly model such data, therefore there is a demand to develop new estimation approaches. To pave the way for modelling such biased circular data, we first introduce

a class of weighted distributions on the circle. We estimate the unknown forms of the distributions in the class by the kernel density estimation method. For posterior predictive density estimation, a Bayesian approach will be outlined and implemented. Numerical assessments, using the MCMC approach, support the findings, via simulation and real data analysis from a Bayesian nonparametric viewpoint.

EO848 Room Virtual R37 COLORED GRAPHICAL MODELS - IN MEMORY OF HELENE MASSAM Chair: Piotr Graczyk

E1626: **Coloured graphical models**
Presenter: **Steffen Lauritzen**, University of Copenhagen, Denmark

An overview of the use of symmetry in Gaussian graphical models will be given. In particular, the connection to group invariance and consider issues of Bayesian inference in such models will be highlighted.

E0703: **Bayesian model selection for colored Gaussian graphic models**
Presenter: **Xin Gao**, York University, Canada
Co-authors: Helene Massam

A class of coloured graphical Gaussian models is considered which is obtained by imposing equality constraints on the precision matrix in a Bayesian framework. The Bayesian prior for precision matrices is given by the coloured G-Wishart which is the Diaconis-Ylvisaker conjugate. We develop a computationally efficient model search algorithm that combines linear regression with a double reversible jump Markov chain Monte Carlo. The latter is to estimate Bayes factors expressed as a posterior probabilities ratio of two competing models. We also establish the asymptotic consistency property of the model determination approach based on Bayes factors. Our procedure avoids an exhaustive search in the space of graphs, which is computationally impossible. Our method is illustrated with simulations and a real-world application with a protein signalling data set.

E0351: **Fused graphical lasso for brain networks with symmetries**
Presenter: **Saverio Ranciati**, Universita di Bologna, Italy
Co-authors: Alberto Roverato, Alessandra Luati

Neuroimaging is the growing area of neuroscience devoted to producing data with the goal of capturing the processes and dynamics of the human brain. We consider the problem of inferring the brain connectivity network from time-dependent functional magnetic resonance imaging (fMRI) scans. To this aim, we propose the symmetric graphical lasso, a penalized likelihood method with a fused type penalty function that takes into explicit account the natural symmetrical structure of the brain. A symmetric graphical lasso allows one to simultaneously learn the network structure and a set of symmetries across the two hemispheres simultaneously. We implement an alternating directions method of multipliers algorithm to solve the corresponding convex optimization problem. Furthermore, we apply our methods to estimate the brain networks of two subjects, one healthy and one affected by mental disorder, and compare them with their symmetric structure. The method applies once the temporal dependence characterizing fMRI data has been accounted for. We compare the impact of the analysis of different detrending techniques on the estimated brain networks. Although we focus on brain networks, the symmetric graphical lasso is a tool that can be more generally applied to learn multiple networks in the context of dependent samples.

E0879: **Bayesian model selection in the space of Gaussian models invariant by permutation symmetry**
Presenter: **Bartosz Kolodziejek**, Politechnika Warszawska, Poland

Multivariate centered Gaussian models for the random variable $X = (X_1, \ldots, X_p)$, invariant under the action of a subgroup of the group of permutations on $\{1, \ldots, p\}$ are considered. Using the representation of the symmetric group on the field of reals, we derive the distribution of the maximum likelihood estimate of the covariance parameter Σ and also the analytic expression of the normalizing constant of the Diaconis-Ylvisaker conjugate prior for the precision parameter $K = \Sigma^{-1}$. We can thus perform Bayesian model selection in the class of complete Gaussian models invariant by the action of a subgroup of the symmetric group, which we could also call complete RCOP models. We illustrate our method with several examples. Further, we present novel results on the normalizing constant of Diaconis-Ylvisaker conjugate prior when X satisfies conditional independencies described by a decomposable graph and the permutation group describing symmetries is a subgroup of the automorphism group of the graph.

E1472: **On normalizing constants of chordal graphical Gaussian models with group symmetry**
Presenter: **Hideyuki Ishi**, Osaka City University, Japan

The graphical Gaussian models are statistical models of central multivariate Gaussian distributions with prescribed conditional independence given by simple graphs. It is known that, if the graph is chordal, then the model admits various explicit calculations. In particular, we have an analytic formula for the normalizing constant of the Wishart distribution of type II, that is, the Diaconis-Ylvisaker conjugate prior for the precision parameter. We consider the graphical Gaussian models invariant under the natural action of a subgroup of the automorphism group of the graph. If the graph is chordal, we obtain an explicit formula for the normalizing constant of the Wishart distributions of type II. In our calculation, some observation of an algebraic structure of a specific block decomposition of the precision matrices is crucial.

EO563 Room Virtual R39 ANALYTICAL ASPECTS WITHIN DEPENDENCE MODELING Chair: Sebastian Fuchs

E1364: **Small and large subclasses of copulas**
Presenter: **Fabrizio Durante**, University of Salento, Italy

In statistical inference for dependence models, most of the procedures are usually proved to work for most underlying copulas, up to an exceptional set that is usually considered "small". However, it is not always natural what a small subset is in this context. For a given metric space, topology offers a natural way of distinguishing small and big sets through Baire categories, although this concept may sometimes be deceptive. We review some previous results about meager and not-meager (i.e. typical) subclasses of bivariate copulas. Moreover, we determine the topological size of some subsets of multivariate quasi-copulas, with particular emphasis on those subsets that are lattice completion of the set of copulas.

E1114: **On weak conditional convergence of bivariate Archimedean and extreme value copulas**
Presenter: **Thimo Kasper**, University of Salzburg, Austria
Co-authors: Sebastian Fuchs, Wolfgang Trutschnig

Looking at bivariate copulas from the perspective of conditional distributions and considering the weak convergence of almost all conditional distributions yields the notion of weak conditional convergence. At first glance, this notion of convergence for copulas might seem far too restrictive to be of any practical importance - in fact, given samples of a copula C the corresponding empirical copulas do not converge weakly conditional to C with probability one in general. Within the class of Archimedean copulas and the class of Extreme Value copulas, however, standard pointwise convergence and weak conditional convergence can even be proved to be equivalent. Moreover, it can be shown that every copula C is the weak conditional limit of a sequence of checkerboard copulas. After proving these three main results and pointing out some consequences some implications for two recently introduced dependence measures and for the nonparametric estimation of Archimedean and Extreme Value copulas are sketched.

E0694: **Dynamical properties of stochastically monotone copulas**
Presenter: **Christopher Strothmann**, TU Dortmund University, Germany
Co-authors: Karl Friedrich Siburg

The behaviour of stochastically monotone copulas with respect to the Markov product of copulas is investigated. These copulas exhibit a unique dependence reduction property, which enables the characterisation of idempotents and n-fold iterates of stochastically monotone copulas as ordinal sums of the independence copula.

E0422: **Copulas products in completely specfied factor models**
Presenter: **Jonathan Ansari**, University of Freiburg, Germany
Co-authors: Ludger Ruschendorf

A completely specified factor model for a risk vector $X = (X_1, ..., X_d)$ is considered where the joint distributions of the components of X with a risk factor Z and the conditional distributions of X given $Z = z$ are specified. We extend the notion of *-product of-copulas as introduced for $d = 2$ and continuous factor distribution previously to the multivariate and discontinuous case. We give a Sklar-type representation theorem for factor models showing that these *-products determine the copula of a completely specified factor model. We investigate in detail the approximation, transformation, and ordering properties of *-products and, based on them, derive general orthant ordering results for completely specified factor models in dependence on their specifications. In particular, we develop tools to derive worst-case distributions in relevant subclasses of completely specified factor models.

E1101: **A measure of multivariate conditional dependence and its estimation**
Presenter: **Wolfgang Trutschnig**, University of Salzburg, Austria
Co-authors: Florian Griessenberger

Working with so-called linkages allows constructing a multivariate, copula-based dependence measure zeta quantifying the strength of dependence of a real-valued (continuous) random variable Y on a d-dimensional (continuous) random vector X. Zeta attains values in [0,1], is 0 if and only if X and Y are independent and 1 if and only if Y is a function of X (in which case we speak of complete dependence of Y on X). Various additional properties of zeta are discussed, as well as related concepts like the underlying equivalent metrics which induce the dependence measure. Moreover, a strongly consistent estimator for zeta which does not require any regularity assumptions on the underlying copula and hence works in full generality is constructed, its speed of convergence is discussed and illustrated in terms of some examples ranging from independence to complete dependence.

EO138 Room Virtual R40 COMPOSITIONAL, DISTRIBUTIONAL AND RELATIVE ABUNDANCE DATA II Chair: Karel Hron

E1161: **Geographically weighted regression analysis with two-factorial compositional covariates**
Presenter: **Kamila Facevicova**, Palacky University Olomouc, Czech Republic
Co-authors: Petra Kynclova, Karel Macku

An approach is introduced to modeling the relationship between a real variable and a relative structure under the presence of spatial dependence. Spatial statistics provides a wide range of methods for the analysis of data with local variations but these methods are rarely accommodated to work with data of a relative nature. The presented methodology is motivated by the problem of modelling local variations of the relationship between at-risk-of-poverty rates and the structure of the German population aged 30-34, given by gender and the highest attained educational level. Consequently, the geographically weighted regression model with covariates forming a compositional table is introduced and interpretation of regression coefficients is discussed.

E1032: **Instrumental variable estimation in compositional regression**
Presenter: **Andrej Srakar**, University of Ljubljana, Slovenia

In an increasing number of empirical studies, the dimensionality (i.e. size of the parameter space) can be very large. In functional data analysis, the appropriate setting to analyze such problems is a functional linear model in which the covariates belong to Hilbert spaces. This has been previously extended to the cases where covariates are endogenous (functional instrumental variables/FIV). This is now extended to the compositional (with extensions also to histogram, i.e. empirical distributional) data setting, with most of the analysis based on compositional regression using additive log-ratio transformation where either or both independent and dependent variables are compositional. We show that estimation leads to an ill-posed inverse problem with a data-dependent operator. We use and extend the notion of instrument strength to compositional and distributional settings and discuss generalized versions of the estimators when the problem is premultiplied by an instrument-dependent operator. We establish appropriate central limit theorems and study the finite sample performance in a Monte Carlo simulation setting. Our application studies the relationship between long term care provision to relatives and paid work, using a recent time use survey from the Survey of Health, Ageing and Retirement in Europe (SHARE). In conclusion, we discuss extensions to other causal inference approaches in the line of distributional synthetic control.

E1081: **Mapping soil texture in the basque country using compositional data analysis: A Bayesian geo-additive approach**
Presenter: **Joaquin Martinez-Minaya**, Basque Center For Applied Mathematics (BCAM), Spain
Co-authors: Dae-Jin Lee, Lore Zumeta-Olaskoaga

Compositional data (CoDa), consisting of proportions or percentages of disjoint categories adding to one, play an important role in many fields such as ecology, geology, etc. The two most popular families to deal with them are the Dirichlet and the logistic-normal using Aitchison geometry. Recent developments on the simplex geometry allow us to express the regression model in terms of coordinates and estimate its coefficients. Once the model is projected in the real space, we can employ a multivariate Gaussian regression to deal with it. In order to allow for more flexibility to relate coordinates with covariates or spatial components, penalized splines smoothing can be employed. One of the main goals is to show how to fit a Geo-additive compositional regression from the Bayesian perspective using the software brms. Another key question when we deal with CoDa is how to perform the model validation process. We propose two Bayesian CoDa regression measures to assess the goodness of fit of the model. This method was applied in mapping soil texture distribution at a finer scale, based on 2279 soil samples surveyed in the Basque Country between 2010-2018. We considered different covariates such as elevation or slope, as well as geological information. The proposed methodology showed a good overall performance in different scenarios providing high-resolution maps at a fine-scale for the agricultural sector in the Basque Country

E1234: **Classification techniques for probability density functions in the Bayes space framework**
Presenter: **Ivana Pavlu**, Palacky University Olomouc, Czech Republic
Co-authors: Karel Hron, Alessandra Menafoglio, Peter Filzmoser, Enea Bongiorno

Classifying observations into one of the pre-existing groups is one of the frequent tasks in mathematical statistics. With the rising availability of functional data, there is a growing demand for suitable methods needed for their proper statistical analysis. When considering probability density functions (PDFs) as a type of functional data, special care should be given to their specific properties, namely scale invariance, relative scale, and possible unit integral representation. In this sense, the Bayes space methodology serves as a framework that enables the use of standard methods of functional data analysis on properly transformed PDFs. Specifically, the centred log-ratio (clr) transformation plays a key role to represent the PDFs in the standard L^2 space. Both parametric (functional logistic regression, functional principal component regression, functional linear discriminant analysis) and nonparametric (k-nearest neighbours algorithm) classification methods are considered, together with a semiparametric approach based on a kernel density estimation. These methods are presented on a geochemical dataset of particle size distributions from four measuring sites in the Czech Republic, serving as natural groups for classification.

E1414: **Isotemporal substitution in time-use behavioural data with compositional scalar-on-function regression**
Presenter: **Paulina Jaskova**, Palacky University Olomouc, Czech Republic

There is a relationship between the proportion of time spent in physical activity (PA) and health. The intensity of PA can be described through an increasing gravitational acceleration of movement. The question is how the relative reallocation of time between PA of various intensities is associated with the health, that is how to find an adequate regression model between the real response and covariate containing the time-use distribution. The time-use relative data can be extended to the continuous case as probability density functions (PDFs). Similar to compositional data, PDFs are also characterized by the scale invariance property and are geometrically represented using the so-called Bayes spaces. After a proper transformation of PDFs to the L_2 space, a standard functional regression model with real response can be used. To describe the effect of certain subintervals of the gravitational acceleration PDF, an isotemporal substitution can be used for amplifying the impact of particular levels of PA on the health outcome. The presented method was applied to a dataset consisting of functional observations resulting from a large study conducted among school-aged children from the Czech Republic. PA variables were modelled as continuous functional outcomes with the percentage body fat mass used as a health indicator. The results confirm the natural expectation as the more time is spent in the PA of higher intensity is associated with a lower percentage of fat mass.

EC856 Room K0.19 (Hybrid 04) CONTRIBUTIONS IN BAYESIAN STATISTICS III Chair: David Rossell

E1441: **Sequential inference for the bayesian mallows model**
Presenter: **Anja Stein**, Lancaster University, United Kingdom
Co-authors: David Leslie, Arnoldo Frigessi

Recently, a Bayesian inference approach has been developed for the Mallows model, a probability distribution that models ranking data. The framework currently uses MCMC methods to learn and sample from the posterior distribution of the model. However, MCMC is computationally costly if the data arrives sequentially, a scenario that can arise in many settings including internet data where individuals express preferences over items and we wish to infer a consensus ranking to inform what to display to future customers, in turn, express preferences. Using MCMC, each time that new data arrives we need to re-run the full MCMC. Instead, we develop a Sequential Monte Carlo (SMC) method to sequentially update our inference with the Bayesian Mallows model. This allows us to efficiently fit the Bayesian Mallows model in scenarios where we receive new observations through time, both in terms of rankings from previously unobserved individuals, and in terms of updated rankings from existing individuals who have previously provided a (partial) ranking. We provide comparison results between the MCMC and SMC approaches for all scenarios considered, using existing MCMC and new SMC code in the BayesMallows R package.

E1518: **A Bayesian hierarchical model for MedDRA coded adverse events in RCTs**
Presenter: **Alma Revers**, Amsterdam University, Netherlands
Co-authors: Michel Hof, Koos Zwinderman

Patients participating in randomized controlled trials (RCTs) often report a wide range of different adverse events (AE) during the trial. MedDRA is a hierarchical standardization terminology to structure the AEs reported in an RCT. The lowest level in the hierarchy is a single medical event, and every higher level is the aggregation of the lower levels. Currently, the AE data of an RCT are often reported as crude rates or exposure-adjusted incidence rates. These rates have limited statistical power to detect rare AEs, leading to a high rate of false negatives. Therefore, we propose a hierarchical Bayesian model for identifying MedDRA coded AEs relative risks (RRs) in an RCT. We started by specifying a hierarchical Binomial and Poisson model, as done by others and extended the models to group the AEs to the complete hierarchy of the MedDRA. We developed our multi-stage hierarchical model, including the complete multiaxial MedDRA structure and developed a Bayesian algorithm to estimate posterior probabilities. We illustrate our model with AE-data from a large RCT ($n = 2658$), and we compare results with other methods for analyzing AEs.

E1658: **A Bayesian test for the equality of two coefficients of variation**
Presenter: **Mara Manca**, University of Cagliari, Italy
Co-authors: Francesco Bertolino, Silvia Columbu, Monica Musio

A Bayesian Discrepancy Test (BDT) for the equality of two coefficients of variation (CVs) concerning independent populations is proposed. It is based on the Bayesian Discrepancy Measure (BDM), a bayesian measure of evidence recently introduced. When taking into account some distributions, such as Lognormal, Poisson, Gamma, Pareto and Weibull, the problem of testing the equality of the CVs of two independent populations is reduced to testing the equality of one of the parameters of the populations considered and the BDT can easily be applied. This simplification does not hold when considering two normal independent populations, but the comparison of their CVs can still be readily managed by using the BDT. The advantage of the proposed procedure is that it does not rely on asymptotic assumptions, unlike most frequentist statistical tests addressed in literature.

E1339: **Heterogeneous treatment effect estimation based on a partially linear model with a Gaussian process prior**
Presenter: **Shunsuke Horii**, Waseda University, Japan

Recently, heterogeneous treatment effect estimation has been attracting a lot of attention due to its importance in various fields. We propose a partially linear model with a Gaussian process prior for the heterogeneous treatment effect estimation. A partially linear model is a semiparametric model that consists of linear and nonparametric components in an additive form. A model that uses a Gaussian process to model the nonparametric component has also been studied in the literature. However, these models cannot handle the heterogeneity of the treatment effect. In the proposed model, not only the nonparametric component of the model, but also the heterogeneous treatment effect of the treatment variable is modeled by a Gaussian process prior. We show the effectiveness of the proposed method through numerical experiments based on synthetic and real-world data.

EG057 Room Virtual R28 CONTRIBUTIONS IN STATISTICAL MODELLING I Chair: Jean-Francois Dupuy

E1668: **Bounded missing data imputation using statistical and machine learning approaches**
Presenter: **Urko Aguirre**, Hospital Galdakao-Usansolo, Osakidetza, Spain
Co-authors: Cruz Borges

Real-life data are mostly bounded and non-Gaussian variables. One of the best approaches for modelling them is the Zero-one-inflated beta (ZOIB) regression. There are no appropriate methods to address the problem of missing data in repeated bounded outcomes. We developed an imputation method using ZOIB (i-ZOIB) and compared its performance with those of the nave and machine-learning methods, using different distribution shapes and settings designed in the simulation study. The performance was measured employing the absolute error (MAE), root-mean-square-error (RMSE) and the unscaled mean bounded relative absolute error (UMBRAE) methods. The results varied depending on the missingness rate and mechanism. There is no consensus among the studied methods: i-ZOIB and the machine-learning ANN, SVR and RF methods showed the best performance.

E0201: **Gamma regression with censored outcomes and missing data**
Presenter: **Jean-Francois Dupuy**, INSA de Rennes, France

Gamma regression is a member of the family of generalized linear models that have proved useful in several domains for modeling a positive-valued outcome as a function of explanatory variables. The case of a right-censored outcome in Gamma regression was recently examined in the literature (right-censoring occurs when the exact outcome is not observed, but is known to be greater than or equal to the observed outcome value). We go a step further and investigate estimation in the Gamma regression model when the outcome is right-censored and censoring indicators are missing at random (MAR). We propose and investigate an augmented inverse probability weighted (AIPW) estimator adapted to this setting. We describe

its asymptotic properties (this estimator is consistent and asymptotically normal) and its double robustness property. We also describe a simulation study that investigates the finite sample performance of the proposed estimate.

E1764: **A proper statistical framework for range-based comparisons of quality attributes**
Presenter: **Gerhard Goessler**, University of Graz, Austria
Co-authors: Vera Hofer, Walter Goessler

When comparing industrial goods (e.g., drug products) with respect to quality attributes (QAs) one often has to deal not only with measurement error but also with a considerable amount of product-related variability. When product-related variability is present, it is not sufficient to just compare mean values - instead, the whole range of possible realizations must be considered. To this end, special statistical tests are needed. Looking, for example, at the respective information published by agencies like the EMA or the FDA, one can see that, unfortunately, such tests together with the proper statistical framework, are at best still under way. Therefore, we give a thorough discussion of the question of when two products can be considered equivalent with respect to a certain property when product-related variability is present. Based on these considerations, we attempt to create a suitable statistical framework for such a comparison which is centered around the novel concept of k-equivalence of two drug products with respect to a certain QA.

E1765: **Statistical tests for range-based comparisons: Increase flexibility for producers without increasing risk for consumers?**
Presenter: **Vera Hofer**, University of Graz, Austria
Co-authors: Gerhard Goessler, Walter Goessler

Based on the concept of k-equivalence some statistical tests for range-based comparisons of quality attributes are discussed. Their properties are compared to statistical approaches proposed by agencies like the FDA or by researchers in academia and in industry. It is demonstrated that most approaches are easy to apply but they are not suited for a range-based comparison of quality attributes. Moreover, advantageous properties of certain statistical approaches are obtained only for a relatively large sample size. The industry can often not utilized them since typically being restricted to a small sample size. Using low numbers of observations may even render statistical comparisons unacceptable owing to the level of patient or producer risk they entail.

E1496: **Copula state-space models with several latent variables**
Presenter: **Ariane Hanebeck**, Technical University of Munich, Germany
Co-authors: Claudia Czado

A state-space model with several latent variables is proposed which uses copulas to get away from the assumption of Gaussian noise. State-space models are an important tool for analyzing time series. They assume that given observations depend on unobserved states over a so-called observation equation. The temporal behavior of the states is explained by the state equation. In many applications, it is assumed that the noise follows a Gaussian distribution. However, this assumption is often not met by real datasets. Copulas are an adequate tool to lift the Gaussian assumption and extend state-space models to very flexible models. For multiple observations with one latent variable, this model was already considered. The natural extension of this model is to allow for more than one latent variable. A Bayesian approach and MCMC-sampling are used to fit the model. Simulated data and real data examples are used to demonstrate the model fit.

EG023 Room Virtual R29 CONTRIBUTIONS IN SPATIAL STATISTICS Chair: Anastassia Baxevani

E0258: **Spatio-temporal cluster detection and disease risk estimation using clustering-based adjacency modelling**
Presenter: **Xueqing Yin**, University of Glasgow, United Kingdom
Co-authors: Gary Napier, Craig Anderson, Duncan Lee

Globally spatially smooth conditional autoregressive (CAR) models are typically used to capture the spatial autocorrelation in areal unit disease count data when estimating the spatio-temporal trends in disease risk. In these models, the spatial autocorrelation structure is typically induced by a binary neighbourhood matrix based on a border sharing specification, such that spatial correlation is always enforced between geographically neighbouring areas. However, enforcing such correlation in the model will mask any discontinuities in the disease risk surface, thus impeding the detection of clusters of areas that exhibit higher or lower risks than their neighbours. Therefore, we propose a novel methodology to account for these discontinuities via a two-stage modelling approach, which either forces the spatial clusters to be the same for all time periods or allows them to evolve dynamically over time. Stage one produces a set of candidate neighbourhood matrices using a variety of common clustering methods that allow for these risk surface discontinuities. Then in stage two, an appropriate spatial autocorrelation structure(s) is selected by estimating the neighbourhood matrix from the candidate set as part of a hierarchical Bayesian spatio-temporal model. The novel methodology is applied to the motivating study of respiratory disease risk in Greater Glasgow, Scotland, from 2011 to 2017.

E1229: **A multi-resolution approximation via linear projection for large spatial datasets**
Presenter: **Toshihiro Hirano**, Kanto Gakuin University, Japan

Recent technical advances in collecting spatial data have been increasing the demand for methods to analyze large spatial datasets. The statistical analysis for these types of datasets can provide useful knowledge in various fields. However, conventional spatial statistical methods, such as maximum likelihood estimation and kriging, are impractically time-consuming for large spatial datasets due to the necessary matrix inversions. To cope with this problem, we propose a multi-resolution approximation via linear projection (M-RA-lp). The M-RA-lp conducts a linear projection approach on each subregion whenever a spatial domain is subdivided, which leads to an approximated covariance function capturing both the large- and small-scale spatial variations. Moreover, we elicit the algorithms for fast computation of the log-likelihood function and predictive distribution with the approximated covariance function obtained by the M-RA-lp. Simulation studies and a real data analysis for air dose rates demonstrate that our proposed M-RA-lp works well relative to the related existing methods.

E1630: **COVID-19 incidence analysis from a spatial functional spectral nonparametric approach**
Presenter: **Felicita Doris Miranda Huaynalaya**, University of Granada, Spain
Co-authors: Maria Dolores Ruiz-Medina

Pure point and continuous spectral approaches are adopted for predicting COVID-19 incidence from a Bayesian and a nonparametric framework, respectively. Firstly, we consider a particular example of the dynamical multiple linear regression model in function spaces. The functional regression parameter vector is estimated in terms of the Bayesian approximation of the functional entries of the inverse covariance matrix operator of the Hilbert-valued error term, by applying generalized least-squares estimation. Under this functional linear modeling, spatial correlations are reflected in the matrix covariance operator of the functional error term. Secondly, we adopt a continuous spectral approach, assuming spatial stationarity in the functional correlation model, representing possible interactions between the COVID-19 incidence curves at the Spanish Communities analyzed. We reformulate, for spatially distributed correlated curves, the nonparametric estimator of the spectral density operator, based on the periodogram operator, in the functional time series context. This estimator allows us to compute the functional regression vector parameter estimator to our spatial functional spectral context. To implement the approach proposed, a computation is developed in the real-data analysis of COVID-19 incidence. Particularly, the non-parametric estimator of the spatial-spectral density kernels, at 1061x1061 cross-times, is computed over the 37x37 spatial nodes of the frequency grid.

E0589: **Interpolation of weather conditions in a flight corridor**
Presenter: **Gong Chen**, Dresden University of Technology, Germany

Co-authors: Hartmut Fricke, Ostap Okhrin, Judith Rosenow

Economic price pressure on airlines requires highly efficient air transport operations. On the other hand, current research initiatives, such as the Single European Sky Air Traffic Management Research Program request a future air traffic system with increased safety, efficiency, and environmental compatibility. Therewith, multi-criteria aircraft trajectory optimization with reliable meteorological information is becoming increasingly important in everyday operations. Data from the Global (Weather) Forecast System are provided at a resolution (28km, 6 hours) that requires interpolation to optimize trajectories with sufficient accuracy (about 200 m, 1 hour). For aerodynamic crucial weather variables such as temperature, wind speed, and wind direction, we investigate different interpolation models such as linear interpolation, Kriging, radial basis function, neural network, and decision tree regression with bagging and boosting. All methods are compared concerning cross-validation interpolation error and computation time. Considering an example trajectory from Prague to Tunis, Monte Carlo simulation is applied to examine how the errors in GFS data and the Kriging interpolation method can have an impact on the simulated trajectory. The results can be used for reliable, in-flight trajectory optimization, where small-scale changes in weather data become highly sensitive input variables.

E1179: **A Bayesian hierarchical spatial copula model**

Presenter: **Mario Martinez Pizarro**, Universidad de Extremadura, Spain

Co-authors: M Isabel Parra Arevalo, Jose Agustin Garcia Garcia, Francisco Javier Acero Diaz

A Bayesian hierarchical framework with a Gaussian copula and a generalized extreme value (GEV) marginal distribution is proposed for the description of spatial dependencies in data. The Bayesian hierarchical model was implemented with a Monte Carlo Markov Chain (MCMC) method that allows the distribution of the model's parameters to be estimated. The results show the GEV distribution's shape parameter to take constant negative values, the location parameter to be altitude dependent, and the scale parameter values to be concentrated around the same value throughout the space. Further, the spatial copula model chosen presents lower deviance information criterion (DIC) values when spatial distributions are assumed for the GEV distribution's location and scale parameters than when the scale parameter is taken to be constant over the whole space.

EG025 Room Virtual R33 CONTRIBUTIONS IN REGRESSION AND REGULARIZATION Chair: Benjamin Poignard

E0344: **Sparse factor models: Asymptotic properties**

Presenter: **Benjamin Poignard**, Osaka University, Japan

Co-authors: Yoshikazu Terada

The problem of estimating a factor model-based variance-covariance matrix is considered when the factor loading matrix is assumed sparse. We develop a penalised Z-estimation framework to handle the identifiability issue of the factor loading matrix while fostering sparsity in potentially all its entries. We prove the oracle property of the penalised Z-estimator for the factor model; that is, the penalisation procedure can recover the true sparse support, and the estimator is asymptotically normally distributed. The non-penalised loss functions are deduced from the class of Bregman divergence losses, providing new estimators for factor modelling. Empirical studies support these theoretical results.

E0522: **Incorporating sparsity, smoothness and group structure in regularized models for spectroscopic data**

Presenter: **Chin Gi Soh**, Nanyang Technological University (National Institute of Education), Singapore

Co-authors: Ying Zhu

High-dimensional spectroscopic data has applications in many fields such as food science, forensic science and biomedical science due to the information it provides about the chemical compositions of the samples. The fitting of classification and regression models to such data is known to be a challenging task due to the high-dimensional setting, as well as the issue of high correlation between spectral variables. One method that has gained interest in recent years is the use of regularization to overcome these challenges. A regularized model for spectroscopic data is presented that incorporates sparsity, smoothness and group structure. The results from some simulation studies on the use of the regularized model will be discussed. An application of the model to Fourier-transform infrared spectroscopic data adulteration studies in olive oil will also be presented. The results suggest that the sparse fused group lasso is able to achieve good prediction performance, while improving on the interpretability of the resulting models.

E1612: **Clusterwise joint lasso with penalty term for discriminating each cluster**

Presenter: **Shinta Urakami**, Doshisha University, Japan

Co-authors: Hiroshi Yadohisa

In analysing high-dimensional data, it is difficult to interpret the results using ordinary multivariate analysis owing to a large number of dimensions. In this case, sparse estimation of the values to be estimated may facilitate the interpretation of the results. In addition, when the data potentially contain a cluster structure, clustering and stratified regression can be used to estimate the regression coefficients for each cluster, after which their coefficients can be interpreted. However, if we apply regression after clustering, we may end up estimating regression coefficients that do not correctly capture the cluster structure of the data. Clusterwise regression analysis, which performs clustering and estimation of regression coefficients simultaneously, may solve this problem. We propose a clusterwise joint lasso with a penalty term for discriminating each cluster. This lasso method simultaneously performs clustering and estimation of regression coefficients and regularizes the objective function to increase the difference in the values of the regression coefficients for each cluster. By applying this method, it is possible to clearly distinguish the values of the regression coefficients among clusters, which may make it easier to identify the characteristics of each cluster.

E0301: **Local machine learning for data giants**

Presenter: **Gilles Cattani**, University of Geneva, Switzerland

Co-authors: Stefan Sperlich, Michael Scholz

Classical nonparametric estimation is the natural link between two cultures: 'traditional regressions methods' and 'pure prediction algorithms'. We borrow ideas of local smoothers and efficient implementation to combine good practices of both cultures to generate a practical tool for the statistical analysis of large data problems, whether estimation, prediction, or attribution. Estimation and prediction are particularly successful when allowing for local adaptiveness. Further, while typically distributed databases are considered as a bane, data localization can turn it into a boon. Similarly, since most of the problems with divide-and-conquer algorithms root in the paradigm of facing a global parameter set, they disappear by localization. Also, the selection of an optimal subsample size is melted with the problem of finding optimal bandwidths, which, moreover, we allow being local too. Finally, model and variable selection can be done and sometimes even becomes necessary when being local. For each step and subprocedure, we looked for the most efficient implementation to keep the procedure fast. The proof of concept and computational details are given in a simulation study. An empirical application to data giants illustrates the practical use of such a tool.

E1757: **Boosting diversity in regression ensembles**

Presenter: **Jean-Michel Poggi**, University Paris-Sud Orsay, France

Co-authors: Jairo Cugliari, Yannig Goude, Mathias Bourel

The practical interest in using ensemble methods has been highlighted in several works. Aggregation estimation, as well as sequential prediction, provide natural frameworks for studying ensemble methods and for adapting such strategies to time series data. Sequential prediction focuses on how to combine by weighting a given set of individual experts while aggregation is mainly interested in how to generate individual experts to improve prediction performance. We look for enhancing these (possibly online) mixture methods by using the concept of diversity. We propose an

algorithm to enrich the set of original individual predictors using a gradient boosting-based method by incorporating a diversity term to guide the gradient boosting iterations. The idea is to progressively generate experts by boosting diversity. Then, we establish a convergence result ensuring that the associated optimization strategy converges to a global optimum. Finally, we show by means of numerical experiments the appropriateness of our procedure using simulated data and real-world electricity demand datasets.

EP002 Room Poster session room I POSTER SESSION (ONLY VIRTUAL) Chair: Elena Fernandez Iglesias

E1306: **Estimation of marginal hazard ratios from observational data under noninformative censoring**
Presenter: **Guilherme Wang de Faria Barros**, Umea University, Sweden
Co-authors: Jenny Haggstrom
When using observational data to study causal relationships, the process of balancing covariates is a critical part of the process to estimate causal effects. Two main distinct approaches are popular when balancing covariates: weighting and matching. In medical and epidemiological studies, one of the most common settings is the time-to-event setting, also known as survival, in which a common causal effect to be estimated is the marginal hazard ratio (MHR). This setting has the particularity of the existence of censoring of the time-to-event, which has been shown to cause bias when estimating causal effects, especially when the censoring mechanism is informative. The goal is to study the properties of weighting and matching for the estimation of the MHR in time-to-event settings under high censoring proportions with noninformative censoring, compare the results of different approaches and suggest possible solutions.

E1319: **Random splitting random forest for survival analysis with non-functional & functional covariates in the EEG-fNIRS trial**
Presenter: **Mohammad Fayaz**, Shahid Beheshti University of Medical Sciences, Iran
Co-authors: Nezhat Shakeri, Alireza Abadi, Soheila Khodakarim
In biostatistics, the survival analysis methods are very popular for analyzing the time-to-event data with censors. The regression tree and random forest are among models that can also handle survival data and we develop a random forest and bagging that can consider the multiple functional covariates with a random-splitting approach. The functional covariate in each tree is randomly split by generating the random number from the exponential distribution and the summary statistics such as average, median, etc. are calculated for these intervals and we put them in the regular algorithm. A new variable importance plot was produced that shows the most important parts of each functional covariate. There are some other functional models such as a functional linear cox regression with Bayesian estimation, optimal estimation, penalized partial likelihood function, joint Bahadur representation of estimators, functional joint models for longitudinal and time-to-event data, functional ensemble survival tree, synthetic data, and additive functional cox models. We applied and compared some of them on a public dataset from the Electroencephalography (EEG) and functional near-infrared spectroscopy (fNIRS) trial for the discrimination/selection response (DSR) task. The functional covariates are the event-related potential of EEG and the time-to-event response is the latency of fNIRS for each group.

E1389: **A one-sided testing procedure for comparing predictive values of two diagnostic tests simultaneously**
Presenter: **Kanae Takahashi**, Hyogo College of Medicine, Japan
Co-authors: Kouji Yamamoto
Diagnostic tests are important for the early detection and treatment of disease in medicine. The positive predictive value (PPV) and the negative predictive value (NPV) describe how good the test is at predicting abnormality, and these are used for quantifying the diagnostic ability of the test. The PPV is the probability of disease when the diagnostic test result is positive, and the NPV is the probability of no disease when the diagnostic test result is negative. There are several methods to compare each of the PPVs and NPVs in paired designs, and a global two-sided testing procedure to simultaneously test the equality of PPVs and NPVs has already been proposed. However, in some cases, it may be necessary to evaluate the superiority of at least one of PPVs and NPVs. We propose a one-sided testing procedure that assesses the superiority of at least one of PPVs and NPVs by applying an approximate likelihood ratio test proposed.

E1486: **On survival trees with competing risks**
Presenter: **Asanao Shimokawa**, Tokyo University of Science, Japan
Co-authors: Etsuo Miyaoka
Although there is extensive research on survival trees, almost all of them assume only one event risk. However, we should consider the mutual effects of various risks (competing risks) if there are several event risks. Therefore, the goal is to construct a tree-structured model that considers the effects of competing risks on the survival time of interesting risks. To achieve this goal, we propose a criterion based on the finite mixture models of the exponential distributions. We evaluate the performance of this criterion through simulation studies. In addition to this, we show the results of the survival tree obtained from actual data.

E1493: **Semiparametric reference curves for EDEN cohort**
Presenter: **Sandie Ferrigno**, INRIA Nancy and University Nancy Lorraine, France
In medicine, reference curves are an important tool for clinical practice. Centile charts for fetal measurements are routinely used to screen for fetuses that may be of abnormal size for gestational age. Data are from the French EDEN mother-child cohort. As measures of child growth, we are studying fetal weight and height that depends on the gestational age, in the third trimester of mothers pregnancy. Semi-parametric methods as LMS are used to construct these curves. In order to compare their performances, AIC and BIC criteria are used.

E1591: **T^2 type test statistic and simultaneous confidence intervals for sub-mean vectors in two-sample problem**
Presenter: **Tamae Kawasaki**, Tokyo University of Science, Japan
Co-authors: Takashi Seo
The test for sub-mean vectors in the two-sample problem is discussed. We consider testing the equality of the mean vectors of the two populations in the first p_1 dimensions (total dimension is p_1+p_2) under the assumption that the last p_2 dimensions in the mean vectors of the two populations are equal. The goal is to propose the T^2 type test statistic and the simultaneous confidence intervals for sub-mean vectors. In order to propose the T^2 type test statistic, we obtain the maximum likelihood estimators, and derive the asymptotic expansion of part of the test statistic for this case where the total sample size is large. We approximate the distribution for the T^2 type test statistics by constant times an F distribution by adjusting the degrees of freedom. The simultaneous confidence intervals for all linear compounds of the difference of two sub-mean vectors are also given. Finally, the accuracy and asymptotic behavior of the approximation are investigated using 1,000,000 trials Monte Carlo simulation.

E1568: **Modelling deprivation indices in Europe using mixture models**
Presenter: **Ivana Mala**, Prague University of Economics and Business, Czech Republic
A Survey of Health, Ageing and Retirement in Europe includes a large spectrum of information on the ageing process of the European population. People of age above 50 are eligible for the survey, and respondents remain in the study until the end of life. In 2013 wave two composite indices of deprivation (material and social) were introduced on the 0-1 scale. The probability distribution of indices in the population is highly country-specific. Mixture models are applied to describe the multimodal distribution and to find out homogenous subgroups of respondents. The Gaussian densities (3-5) are complemented with a positive probability of no material deprivation. The estimated component distributions are used to compare the situation in participating countries and specify countries with similar distributions. The mixture model enables us to compare level, variability and component frequencies and the number of components needed to describe the distributions. Moreover, the differences between distributions

of both indices are shown. The problem of social deprivation of the elderly population, even in welfare states of Europe, is also discussed and quantified based on these indices.

C1555: **Forecasting under breaks in non-parametric regression**
Presenter: **Sze Him Isaac Leung**, The Chinese University of Hong Kong, Hong Kong
Co-authors: Chun Yip Yau

The prediction of future observations in a non-parametric regression model which is subject to a structural break in time is studied. We propose a weighted kernel estimator to estimate the post-break function, where the weights are both time and location dependent. It is shown that incorporating pre-break observations can improve the estimate of the post-break 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. Simulation studies indicate that the proposed weighted kernel estimator has a lower MSFE compared with traditional post-break methods.

C1593: **Deep reinforcement learning in portfolio selection**
Presenter: **Lenka Nechvatalova**, Charles University, Czech Republic
Co-authors: Jozef Barunik

Reinforcement learning is used to form portfolios for investors with asymmetrical and distorted utility functions. These utility functions do not allow finding optimal portfolio weights as an analytical or straightforward optimization solution. Reinforcement learning is a class of machine learning algorithms where an agent with the goal of maximizing a long-term reward is sequentially making decisions while interacting with the environment and learning from her experience. The portfolio formation is demonstrated on a number of theoretical examples using simulations as well as on empirical datasets. The resulting portfolios are compared with portfolios formed using traditional portfolio selection methods.

E1775: **Extreme flood frequency trends inferential analysis in Nalon River (Asturias, NW Spain)**
Presenter: **Elena Fernandez Iglesias**, University of Oviedo, Spain
Co-authors: Gil Gonzalez-Rodriguez, Veronica Moro Garcia

Economic losses due to the effect of river floods are one of the largest natural hazard costs in the world. Those are expected to increase due to the growth of population and, according to different studies, to climate change. Several researches show that climate change may alter the frequency and magnitude of river floods in Europe, including Spain. Due to the lack of long records of maximum flows only a few studies at regional scales are available. In order to analyse the flood trends in the Nalon river, the biggest fluvial basin in Northwest of Spain, flow records are extended with other kind of historical flood data to obtain a time series of extreme events, with a return period over 25 years to the period 1938–2020. A simple descriptive analysis shows fairly symmetric distributions for the time between extreme events. In addition, the extreme floods events frequency has been doubled since 2000. A Bootstrap isotonic test in strict sense was applied showing that, indeed, there is an effective increase of the frequency of extreme floods.

C1778: **Credit creation and risk in opaque intermediaries**
Presenter: **Charis Eleftheriou**, Cyprus University of Technology, Cyprus
Co-authors: Panayiotis Andreou, Demetris Koursaros

Banks are believed to be opaque to outsiders, since opacity is a property of the assets that they hold. A social planner would like banks to be transparent, riskless and highly efficient intermediators of liquidity. However, these goals appear to be conflicting. Whether and how opacity and intermediation are connected is an important question of relevance to regulators, investors and the general public. The theoretical literature makes conflicting predictions about this relation. We formulate a theoretical model of opacity and intermediation and test the resulting conjectures on a large sample of US banks. We find that intermediation is positively associated with opacity, while controlling for a large number of other factors including fragility. These results imply that demanding full disclosure and transparency from banks may bring with it negative externalities in terms of intermediation, which policymakers may wish to take into account.

CO266 Room K E. Safra (Multi-use 01) NONLINEAR AND FINANCIAL TIME SERIES (HYBRID) Chair: Ruijun Bu

C0316: **On the right jump tail inferred from the VIX markets**
Presenter: **Xingzhi Yao**, Xián Jiaotong Liverpool University, China
Co-authors: Zhenxiong Li, Marwan Izzeldin

Using the VIX futures and options data from 2006 through 2020, it is shown that the right jump tail implied by the VIX out-of-the-money call options is a significant risk factor that affects VIX option returns, beyond the volatility of volatility (VOV) and implied skewness as documented in the recent literature as relevant predictors. A comprehensive simulation study reveals the non-trivial effects on the dynamics and term structure of VOV, implied skewness and variance-of-variance risk premium of upward jumps in the VIX. In addition, strong evidence is delivered in support of the hypothesis that the right jump tail is the key driver of the VOV and implied skewness and fully subsumes the information contained in the left jump tail.

C0399: **Sequential monitoring for change points of M-estimators in risk models**
Presenter: **Xiaohan Xue**, ICMA Centre, University of Reading, United Kingdom

A new real-time detection method is proposed for change points of M-estimators in a risk model, based on a previous study. The proposed test efficiently captures change points in several risk measure series: volatility, expectile, Value-at-Risk (VaR), and Expected Shortfall (ES). We derive the asymptotic distribution of the proposed statistic. Monte Carlo simulation results show that our proposed test has better size control and higher power under various change point scenarios. The empirical studies of risk measures based on the S&P 500 index and GBP/EUR exchange rate illustrate that our proposed test is able to detect change points that are consistent with well-known market events.

C0482: **Time-varying spillover networks of green bond and related financial markets**
Presenter: **Xiaohang Ren**, Central South University, China
Co-authors: Ping Wei, Kang Yuan

Using the Granger causality test and spillovers network analysis based on the time-varying parameter vector autoregressive (TVP-VAR) model, the interrelationship between the green bond market and other major financial markets are investigated. The empirical findings reveal: (i) there exists a significant bidirectional spillover effect between the green bond market and the U.S. Treasury market; (ii) during periods of economic turmoil, the connectedness between green bonds and other markets has increased significantly, especially the risk spillovers from the stock markets; (iii) The carbon market and the energy futures market only had spillovers on the green bond market before the publication of the Green Bond Principles in 2014. We discuss the influence of COVID-19 on the spillovers network. These findings have strong implications for investors to manage portfolios and policymakers to improve the regulations.

C1040: **An econometric analysis of regression models with sorted variables**
Presenter: **Sha Meng**, University of Liverpool, United Kingdom

The quintile portfolio-level analysis (QPA) is widely used in the empirical finance literature to investigate the predictive power of a tested risk factor for assets returns. The QPA assumes that the individual stock return is a linear combination of controlling risk factors and tested risk factors. The hypothesis testing uses the cross-section of stock returns to create stock portfolios that have different sensitivities to the tested risk factor and

estimates regressions for sorted portfolios and then calculates T statistics. However, very few studies have investigated the econometric properties and finite sample performance of QPA test. A model framework is constructed for the data generating process and analyses the size and power of QPA. Also, the standard QPA procedure is modified and a direct test procedure is proposed to directly test the assumption in panel data regressions. The size and size-corrected power of QPA and direct test are compared using the same datasets, and results show that sizes of the two tests are similar and there are no size distortions, whereas, the size-adjusted power of QPA test increases more slowly comparing with that of the direct test, which means that the direct test is much more powerful than QPA.

C1293: **Modelling endogenous regime-switching: A copula-based latent factor approach**
Presenter: **Yuyi Li**, University of Liverpool, United Kingdom
Co-authors: Ruijun Bu, Jie Cheng

A copula-based latent factor-driven endogenous regime-switching model is proposed with a general dependence structure between the innovation of the regime-dependent process and that of the latent factor assumed to be driving the switching of the regimes. In the spirit of copula modelling, we first specify the marginal distribution of regime-dependent innovation and that of the latent factor independently, and then model their dependence either by a copula function or by directly specifying their conditional behaviour. Consequently, our model can accommodate potentially non-Gaussian regime-dependent dynamics and nonlinear endogenous dependence, which are typically observed in empirical data. Similar to existing models in the literature, the parameters of the proposed model can be estimated by maximum likelihood using a modified Markov switching filter and the unobservable latent factor can be extracted using a smoothing technique. The flexibility and usefulness of our model are illustrated by numerical examples.

CO551 Room K0.16 (Hybrid 02) CONTRIBUTIONS IN COMMODITY MARKETS AND ASSET PRICING Chair: Francesco Poli

C1774: **Investor attention spillover between equity market and commodity market**
Presenter: **Nan Zhao**, Cass Business School, City, University of London, United Kingdom
Co-authors: Ana-Maria Fuertes

Exploiting the measure of investors' attention from media coverage of news articles, we find that commodity futures with related stocks that experience higher returns in the past two weeks are associated with higher returns and turnover in the future week, after adjusting for a battery of risk and characteristic benchmarks. This finding is consistent with our conjectures that investors (a) tend to trade more after a positive investment experience, and (b) are more likely to transfer their attention from the stock market to the commodity market.

C1169: **Marine fuel hedging under the sulphur cap regulations**
Presenter: **Frantisek Cech**, UTIA AV CR, v.v.i., Czech Republic

It is argued that consumers and producers of marine fuels can reduce the uncertainty of their portfolios under the environmental regulations aimed at air pollution reduction. The results show that variance reduction can be up to 72% compared to unhedged position. We also identify Gasoil futures as the universal hedging instrument to manage uncertainty.

C1781: **Contagion in commodities markets**
Presenter: **Olusiji Sanya**, Birkbeck College, University of London, United Kingdom
Co-authors: Roald Versteeg, Emanuela Sciubba

Contagion is investigated within the exchange-traded commodities complex (US Dollars denominated indices for markets within the agriculture, energy, industrial metals and precious metals groupings between January 1999 to February 2013) resulting from the 2008 commodities market crash using a system of non-linear simultaneous equations which considers both bad and good contagion, resulting from extreme negative shocks and extreme positive shocks respectively. The methodology also empowers us to endogenise crisis periods and investigate contagion to and from multiple sources. Separating interdependence and monsoon effects, we find industrial metals and energy to be the most systemically important sources of contagion. We capture reverse contagion effects, i.e, the crisis triggered a flight to quality effects captured through counterintuitive, but statistically significant, contagion indicators which fundamentals cannot explain. The results also empirically corroborate the financialization of exchange-traded commodities markets as debated in the literature.

C1402: **Pricing risk in cryptocurrency market: A (mis)calculated gamble?**
Presenter: **Jan Sila**, Univerzita Karlova, Czech Republic
Co-authors: Ladislav Kristoufek

The pricing of risk in the cross-section of cryptocurrency returns is inspected. We present new empirical evidence concerning systemic risk and consider the implications for asset pricing through portfolio sorts. It is shown that, in fact, the risk premium in the market is priced and is positive. Thus, participants pay for the exposure to systemic risk. There is evidence that it has changed in the recent bull run, however the main results this holds. We study a parsimonious model of the return on the market and systemic risk factors from 2015 to 2021. Due to the rapidly changing universe of cryptocurrencies, we take a snapshot of the market at the end of each month and select candidates only from the then relevant coins. At the beginning of the sample period, we demonstrate that, contrary to the standard stocks-related literature, investors seek exposure to the systemic risk, as currencies with higher risk factor loadings earn higher returns. We argue that this has turned around in the recent surge which suggests that the market has matured from a seemingly gambling phase into a risk-rational investment one.

C1639: **Idiosyncratic quantile risk and asset prices**
Presenter: **Matej Nevrla**, UTIA AV CR vvi, Czech Republic
Co-authors: Jozef Barunik

The aim is to investigate common movements of idiosyncratic components of asset returns in their quantiles. Using a large panel of stock returns, we estimate latent idiosyncratic quantile factors, define and estimate corresponding betas, and analyze their significance for both time-series and cross-sectional implications. We specifically focus on the fluctuations in the left tail of the return distributions. To estimate the quantile factors, we employ state-of-the-art methods and select the appropriate number of factors sufficient to describe the distributions' common movement. The added value is assessed with respect to the idiosyncratic volatility risk and other tail risk measures widely used in the literature. The results are of importance for both theoretical asset pricing models and empirical efforts in the literature.

CO808 Room K0.18 (Hybrid 03) ADVANCES IN VOLATILITY MODELING (VIRTUAL) Chair: Christian Francq

C0172: **Identifying structural shocks to volatility through a proxy-MGARCH model**
Presenter: **Jeannine Polivka**, University of St.Gallen, Switzerland
Co-authors: Matthias Fengler

The classical MGARCH specification for volatility modeling is extended by developing a structural MGARCH model targeting the identification of shocks and volatility spillovers in a speculative return system. Similarly to the proxy-sVAR framework, we consider auxiliary proxy variables constructed from news-related measures to identify the underlying shock system. We achieve full identification with multiple proxies by chaining Givens rotations. In an empirical application, we identify an equity, bond and currency shock. We study the volatility spillovers implied by these labelled structural shocks. Our analysis shows that symmetric spillover regimes are rejected.

C0584: **Testing the existence of moments and estimating the tail index of augmented GARCH processes**
Presenter: **Christian Francq**, CREST and University Lille III, France
Co-authors: Jean-Michel Zakoian

The purpose is to investigate the problem of testing finiteness of moments for a class of semi-parametric augmented GARCH(1,1) models encompassing the most commonly used GARCH-type specifications. The existence of positive-power moments of the strictly stationary solution is characterized through the Moment Generating Function (MGF) of the model, defined as the MGF of the logarithm of the random autoregressive coefficient in the volatility dynamics. We establish the asymptotic distribution of the empirical MGF, from which tests of moments are deduced. Alternative tests relying on the estimation of the maximal exponent characterizing the existence of moments are studied. The fully parametric case where the innovations density is either known or estimated is also considered. Power comparisons based on local alternatives and the Bahadur approach are proposed. Our results are illustrated via Monte Carlo experiments and real financial data.

C0607: **Testing hypotheses on the innovations distribution in GARCH-type models**
Presenter: **Jean-Michel Zakoian**, CREST, France
Co-authors: Christian Francq

Tests of different hypotheses in general GARCH models are proposed: adequacy of a parametric quantile, mean-median equality, symmetry of extreme quantiles and zero-median in presence of a conditional mean. The tests rely on the asymptotic distribution of the empirical distribution function of the residuals (edfr). For a large class of time series models (including the standard ARMA-GARCH),the asymptotic distribution of the edfr is impacted by the estimation but does not depend on the model parameters. The resulting tests are generally model-free (though not estimation-free) and thus are simple to implement. Efficiency comparisons are made using the Bahadur approach. A numerical study based on simulated and real data is provided.

C0700: **A multivariate ARCH(∞) model with exogenous variables and dynamic conditional betas**
Presenter: **Julien Royer**, CREST, France
Co-authors: Jean-Michel Zakoian, Christian Francq

Factor models are highly common in the financial literature. Recent advances allow relaxing the constancy of slope coefficients (the so-called betas) by considering conditional regressions. The theory on the estimation of these dynamic conditional betas however usually relies on short memory volatility models, which can be restrictive in empirical applications. Moreover, exogenous variables have proven useful in recent studies on volatility modeling. We introduce a multivariate framework allowing for time-varying betas in which covolatilities can exhibit higher persistence than the standard exponential decay. Covariates are included in the dynamics of both conditional variances and betas. We establish stationarity conditions for the proposed model and prove the consistency and asymptotic normality of the QML estimator. Monte Carlo experiments are conducted to assess the performance of the estimation procedure in finite sample. Finally, we discuss the choice of potential relevant exogenous variables and illustrate the pertinence of the model on real data applications.

C0754: **Modelling volatility cycles**
Presenter: **Christian Conrad**, Heidelberg University, Germany
Co-authors: Robert Engle

A multiplicative factor multi-frequency component GARCH model is proposed which exploits the empirical fact that the daily standardized forecast errors of standard GARCH models behave counter-cyclical when averaged at a lower frequency. For the new model, we derive the unconditional variance of the returns, the news impact function and multi-step-ahead volatility forecasts. We apply the model to more than 5,000 assets. We show that the long-term component of stock market volatility is driven by news about the macroeconomic outlook and monetary policy as well as policy-related news. The new component model significantly outperforms the nested one-component GJR-GARCH and several HAR-type models in terms of out-of-sample forecasting.

CO633 Room K0.20 (Hybrid 05) FINANCIAL ECONOMETRICS: MODELLING AND FORECASTING (VIRTUAL) Chair: Vincenzo Candila

C0727: **Forecasting VaR and ES using a joint quantile regression and its implications in portfolio allocation**
Presenter: **Luca Merlo**, Sapienza University of Rome, Italy
Co-authors: Lea Petrella, Valentina Raponi

A multivariate quantile regression framework is proposed to forecast Value at Risk (VaR) and Expected Shortfall (ES) of multiple financial assets simultaneously. We generalize the Multivariate Asymmetric Laplace (MAL) joint quantile regression to a time-varying setting, which allows us to specify a dynamic process for the evolution of both the VaR and ES of each asset. The proposed methodology accounts for the dependence structure among asset returns. By exploiting the properties of the MAL distribution, we propose a new portfolio optimization method that minimizes portfolio risk and controls for well-known characteristics of financial data. We evaluate the advantages of the proposed approach on both simulated and real data, using weekly returns on three major stock market indices. We show that our method outperforms other existing models and provides more accurate risk measure forecasts than univariate methods.

C0860: **Time-varying Poisson autoregression**
Presenter: **Giovanni Angelini**, University of Bologna, Italy

A new time-varying econometric model is proposed, called Time-Varying Poisson AutoRegressive (TV-PAR), with is suited to forecast time series of counts. We show that the score-driven framework is particularly suitable to recover the evolution of time-varying parameters and provides the required flexibility to model and forecast time series of counts characterized by convoluted nonlinear dynamics and structural breaks. We study the asymptotic properties of the TV-PAR model and prove that, provided some mild conditions, maximum likelihood estimation (MLE) yields strongly consistent and asymptotically normal parameter estimates. Finite-sample performance and forecasting accuracy are evaluated through Monte Carlo simulations. The empirical usefulness of the time-varying specification of the proposed TV-PAR model is shown by analyzing the number of monthly corporate defaults.

C1036: **How do autoregressive processes help forecasting the GDP**
Presenter: **Giovanni Maccarrone**, La Sapienza, Italy
Co-authors: Giacomo Morelli, Sara Spadaccini

The predictive power of different models to forecast the real U.S. GDP is compared. Using quarterly data from 1976 to 2020, we find that the machine learning K-Nearest Neighbour (KNN) model captures the self-predictive ability of the U.S. GDP and performs better than traditional time series analysis. We explore the inclusion of predictors such as the yield curve, its latent factors, and a set of macroeconomic variables in order to increase the level of forecasting accuracy. The predictions result to be improved only when considering long forecast horizons. The use of machine learning algorithms provides additional guidance for data-driven decision making.

C1236: **Do fiscal policies affect the firms growth and performance? Evidence from a regional study in the Dominican republic**
Presenter: **Marinella Boccia**, University of Salerno, Italy
Co-authors: Alessandra Amendola, Gianluca Mele, Luca Sensini

The effect of corporate income tax (CIT) on business performance in the Dominican Republic is investigated, focusing attention on both the national and regional levels. The analysis is based on data provided by the local authorities of the Dominican Republic to the World Bank and includes

administrative CIT declarations for the period 2006-2015 of over 18,000 companies distributed in 31 provinces. A propensity score matching method is carried out considering some opportunely selected financial indicators as proxies of firms performance. The overall results show that CIT incentives have a positive impact on growth and on most performance indicators, however pointing out some significant differences between the regions.

C1254: **Forecasting the production of renewable energy**
Presenter: **Sara Spadaccini**, La Sapienza, Italy
Co-authors: Giovanni Maccarrone, Giacomo Morelli

Renewable energy production is forecasted in European countries relying on several factors that potentially affect its behavior. We identify common factors across and peculiarities within countries. We employ different statistical techniques to detect the determinants that have a relevant impact on the renewable energy market. The benefit for a policymaker is to understand which factors and policies contribute to easing the renewable energy transition.

CO218 Room Virtual R35 OPTIMIZATION MODELLING IN STRUCTURAL ECONOMETRICS Chair: Stefano Nasini

C0385: **Endogenous learning in input-output economies**
Presenter: **Stefano Nasini**, IESEG School of Management, France
Co-authors: Nessah Rabia

Consider a multisector general equilibrium model where firms have incomplete information about the return to scale of their production and that information is sequentially updated once real production is observed. What is the impact of these learning dynamics on the market-wise equilibrium objects? Under which conditions firms are able to efficiently learn their actual return to scale? At which rate does this learning happen? We analyze endogenous learning mechanisms and their implications for the market-wise equilibrium objects in the multisector model. The results shed light on how idiosyncratic shocks translate into learning dynamics of the input-output elasticity structure. In particular, we observe that (i) all the relevant information in the learning dynamics are encoded in the input decisions; (ii) firms are able to learn the actual return to scale independently from the way in which input decisions are taken; (iii) the mismatch between the true (unknown) returns to scale and the ones predicted by firms has a critical effect on the aggregate production, which is amplified when the economy is intensive in capital.

C0386: **An inferential approach for the influential-imitator diffusion**
Presenter: **Ringo Thomas Tchouya**, IMSP, Benin
Co-authors: Stefano Nasini, Sophie Dabo

A new inferential approach is proposed for influential-imitator dynamics, a widely accepted extension of the traditional Bass diffusion model, for cases where the population has two segments: influentials (who influence each other) and imitators (whose choices are affected by those of the influentials). The main difficulty in using this continuous-time diffusion model is that the solution of the underlying differential equation is not an explicit function of its unknown parameters. We have some main results in the context of estimating the parameters of the influential-imitator dynamics. (1) We develop a truncated power series, providing an explicit solution of the differential equation; this results in an asymptotically correct approximation, with increasing accuracy as the spontaneous innovation/adoption parameter decreases. (2) we show that a block decomposition of the underlying parameter matrix leads to a double truncation of the power series which allows expressing explicitly the dependence of the likelihood function on the unknown parameter. After a detailed analysis of the theoretical properties, the proposed estimation approach is empirically tested using Michell and West's dataset of cannabis use by a cohort of students during their second, third and fourth year at a Glasgow high school.

C0496: **Nash bargaining partitioning in decentralized portfolio management**
Presenter: **Nessah Rabia**, IESEG School of Management, France
Co-authors: Francisco Benita, Stefano Nasini

In the context of decentralized portfolio management, understanding how to distribute a fixed budget among decentralized intermediaries is a relevant question for financial investors. We consider the Nash bargaining partitioning for a class of decentralized investment problems, where intermediaries are in charge of the portfolio construction in heterogeneous local markets and act as risk/disutility minimizers. We propose a reformulation that is valid within a class of risk/disutility measures (that we call quasi-homogeneous measures) and allows the reduction of a complex bilevel optimization model to a convex separable knapsack problem. As numerically shown using stock returns data from U.S. listed enterprises, this modelling reduction of the Nash bargaining solution in decentralized investment (driven by the notion of quasi-homogeneous measures), allows solving the vast majority of large-scale investment instances in less than a minute.

C0526: **Large-scale smart evaluation of risk attitude: A structural econometric framework**
Presenter: **Nathalie Picard**, University of Strasbourg, France
Co-authors: Andre de Palma, Stefano Nasini

A statistical methodology is built to put together decision-making under risk and uncertainty (non-expected utility, behavioral economics and discrete choice theory), structural econometrics (hierarchical modelling, limited dependent variables and latent variables, Bayesian posterior) and optimization methods for large-scale estimation (Alternating Direction methods, and Specialized Markov Chain Monte Carlo methods). The main scope is to estimate and disentangle the different dimensions of investors risk attitude (risk aversion, loss aversion and probability weighting). To numerically assess and empirically validate the correctness and efficiency of our methodology, we use investors data and data collected via the online multilingual questionnairehttps://riskdynametrics.com since 2004.

C0527: **Economic distributions, primitive distributions, and demand recovery in monopolistic competition**
Presenter: **Andre de Palma**, CY Cergy-Paris UniversitÃl', France
Co-authors: Simon Anderson

Fundamental technological and taste distributions are linked to endogenous economic distributions of prices and firm size (output, profit) generated under monopolistic competition with heterogeneous productivities as per recent Trade and IO models. We derive new properties for monopoly pricing and equivalence properties on demand curvature, profit functions, and marginal revenue, which we use to ensure distributions of cost, price, output, and profit can be matched under monopolistic competition. Demand and one distribution determine the rest. We provide constructive proofs to recover demand and all distributions from just two (e.g., price and cost distributions uncover demand form), and derive consistency conditions that distribution pairs must satisfy. We then extend to include mark-up distributions

CO034 Room Virtual R38 TOPICS IN FINANCIAL ECONOMETRICS Chair: Leopold Soegner

C1186: **Bayesian reconciliation of the return predictability**
Presenter: **Borys Koval**, Vienna University of Economics and Business, Austria
Co-authors: Sylvia Fruehwirth-Schnatter, Leopold Soegner

A Vector Autoregression (VAR) for the returns, dividend growth and the dividend-price ratio is estimated, where the Bayesian Control Function approach is applied to account for biased coefficient estimates in the predictive regression. Motivated by financial literature we impose a stationarity condition on the auto-regressive dividend-price ratio process employing Bayesian priors. We develop two new priors, a Jeffreys prior and a prior

based on the frequentist reduced-bias approach, and compare our Bayesian estimation routine to other approaches proposed in the literature (e.g., uniform and Reference prior) by means of an extensive simulation study. In terms of mean squared error (MSE), mean absolute error (MAE), and credible interval coverage, the proposed approach leads to superior performance relative to ordinary least squares estimation, a frequentist reduced-bias approach, and Bayesian estimation using priors proposed in the literature. We apply our methodology to financial data for the S&P 500 and find strong evidence for return predictability after properly accounting for the correlation structure and imposing theory-motivated restrictions on the dividend-price ratio.

C1191: **Conditional skewness in currency markets**
Presenter: **Alina Steshkova**, Vienna University of Economics and Business, Austria

The aim is to introduce the modeling of currency returns conditional on the interest rate differential and the real exchange rate under the assumption that FX returns are skew-t distributed. Beyond the well-known relationship between the currency risk premium and its risk factors, we document a negative effect of the interest rate differential and the real exchange rate on idiosyncratic skewness in a short-term horizon for a set of developed and emerging market currencies. The out-of-sample results reveal evidence of the predictability of idiosyncratic skewness and the ability of conditional skewness to forecast currency risk premia at the country level. The risk factor based on conditional skewness is priced across carry trade portfolios and improves the understanding of the carry trade strategy returns. Flexible modeling of the return asymmetry and fat tails provides an accurate forecast of the value-at-risk and expected shortfall, and contributes to the analysis of downside risk in the FX market.

C1329: **Forecast combinations for benchmarks of long-term stock returns using machine learning methods**
Presenter: **Michael Scholz**, University of Klagenfurt, Austria

Forecast combinations are a popular way of reducing the mean squared forecast error when multiple candidate models for a target variable are available. We apply different approaches to finding (optimal) weights for forecasts of stock returns in excess of different benchmarks. The focus lies thereby on nonlinear predictive functions estimated by a fully nonparametric smoother with the covariates and the smoothing parameters chosen by cross-validation. Based on an out-of-sample study, we find that individual nonparametric models outperform their forecast combinations. The latter are prone to in-sample over-fitting and in consequence, perform poorly out-of-sample especially when the set of possible candidates for combinations is large. A reduction to one-dimensional models balances in-sample and out-of-sample performance.

C1661: **REMIS (Retrieval from Mixed Sampling Frequency) for VAR(MA)s**
Presenter: **Philipp Gersing**, Vienna University of Technology, Austria
Co-authors: Manfred Deistler

A novel estimation technique for a general approach to mixed frequency data is proposed which we label REtrieval from MIxed Sampling Frequency data (REMIS): Given data observed at mixed sampling frequency, we retrieve the parameters of an underlying high frequency VAR(MA) system. A canonical state space representation is derived for the blocked process which second moments correspond to all those second moments observable under mixed frequency. Based on this representation, we derive a consistent estimator for the interpolation of the missing second moments. The high-frequency parameters are then obtained by Yule Walker estimation with the full set of second moments. We demonstrate the effectiveness of our method compared to other techniques in a simulation study and on an empirical application for GDP imputation and forecasting.

C1637: **REMIS (Retrieval from Mixed Sampling Frequency) for generalised dynamic factor models**
Presenter: **Christoph Rust**, WU Vienna University of Economics and Business and University of Regensburg, Germany
Co-authors: Philipp Gersing, Manfred Deistler, Leopold Soegner

Representation and estimation theory for Generalised Dynamic Factor Models with mixed frequency data is provided. We suppose a GDFM for the underlying high-frequency processes where the spectrum of the common component is rational. We look at the structure of the stacked process running on the slow frequency sampling rate containing all observable outputs. With this approach, we aim to build "information efficient" methods for denoising, parameter estimation and factor extraction with observations under mixed sampling frequency. We prove that the blocked process is again a GDFM with rational spectrum in the common component and define a canonical state-space realisation for the blocked common component. We analyse the relationship between the dynamic and static factor spaces of the underlying high- and low-frequency processes and the blocked process and show under which conditions the high-frequency dynamic factor space can be retrieved from mixed frequency data. Besides forecasting, imputation of the slow/aggregated variables is also a possible application of our results. We demonstrate our methods in an empirical application on forecasting macroeconomic variables with a high dimensional panel of quarterly and monthly observed US-macroeconomic time series.

CO038 Room K2.31 Nash (Hybrid 07) STRUCTURAL SHOCKS AND THEIR PROPAGATION Chair: Rustam Ibragimov

C0780: **Do market-based network reflect true exposures between banks?**
Presenter: **Madina Karamysheva**, Higher School of Economics, Russia
Co-authors: Ben Craig, Dilyara Salakhova

Due to the lack of or poor access to the data on real exposures between banks, several methods have been proposed to reconstruct a network using market data. However, what does this market-based network represent? We replicate several well-known methods to construct market-based networks. Next, we build networks based on true exposures through loans and securities holdings. Then we provide graphical analysis as well as a comparison of network characteristics across different types of networks and different time periods. The regression analysis sheds light on which balance-sheet exposures better explain the links perceived by the market. The findings suggest that while global network structure remains stable, networks evolve over time. Regression analysis shows that the market identifies two banks as connected when they have similar business models defined by overlapping portfolios more than exposures through lending or securities holdings.

C0832: **Credit supply shocks and household defaults**
Presenter: **Anna Pestova**, CERGE-EI, Czech Republic
Co-authors: Mikhail Mamonov

Are disruptions of the mortgage market a consequence of financial imbalances accumulated in the past? We study the effects of positive and negative credit supply (CS) shocks on subsequent household defaults on debt over the last four decades in U.S. states. We apply sign restrictions within a VAR framework to isolate state-level CS shocks, and identify that 1984 and 2004 were the years of systemic, countrywide, positive CS shocks whereas 1989 and 2009 brought systemic negative shocks. Further, by employing a difference-in-differences framework, we find that both positive and negative CS shocks lead to greater household defaults in the future if they also increase mortgage-to-income ratios. We show that the CS shock-induced (i) shifts of employment between the tradable and non-tradable sectors, (ii) changes in household income and (iii) in house prices facilitate the accumulation of default risks. Our results indicate that positive CS shocks occurred in 1984 did not raise household defaults by more in more exposed states compared to less exposed states because the shocks increased both future income and mortgage debt, while not affecting mortgage-to-income ratios. In contrast, the 1989, 2004 and 2009 CS shocks increased mortgage-to-income ratios in subsequent years, thereby raising debt delinquencies and household defaults. These results provide further empirical evidence to theories of endogenous credit cycles.

C1116: **On robust testing for trend**
Presenter: **Anton Skrobotov**, Russian Presidential Academy of National Economy and Public Administration and SPBU, Russia

A simple approach is provided for robust testing for the trend function in the time series under uncertainty over the order of integration of the

error term. The proposed approach relies on the asymptotic normality of the trend coefficient estimator and utilises a t-statistic approach based on splitting the sample. The Monte-Carlo results demonstrate that the approach has the correct finite sample size and favorable finite sample power properties for all data generating processes considered. The proposed approach is robust to very general assumptions on the error term including various forms of non-stationary volatility and heavy tails.

C1716: **COVID-19: Tail risk and predictive regressions**
Presenter: **Rustam Ibragimov**, Imperial College London and St. Petersburg State University, United Kingdom
Co-authors: Walter Distaso, Alexander Semenov, Anton Skrobotov

The focus is on an econometrically justified robust analysis of the effects of the COVID-19 pandemic on financial markets in different countries across the World. It provides the results of robust estimation and inference on predictive regressions for returns on major stock indexes in 23 countries in North and South America, Europe, and Asia incorporating the time series of reported infections and deaths from COVID-19. We also present a detailed study of persistence, heavy-tailedness and tail risk properties of the time series of the COVID-19 infections and death rates that motivate the necessity in applications of robust inference methods in the analysis. Econometrically justified analysis is based on heteroskedasticity and autocorrelation consistent (HAC) inference methods, recently developed robust-statistic inference approaches and robust tail index estimation.

C1398: **Time series models for tracking and forecasting epidemics**
Presenter: **Paul Kattuman**, University of Cambridge, United Kingdom
Co-authors: Andrew Harvey

As an epidemic takes hold, projections of its trajectory enable health care providers to plan and organize clinical and human resources to meet treatment requirements. A new class of time series models is developed that reflect epidemic trajectories and are able to make good forecasts even before new cases and/or deaths reach their peak. The models are relatively simple and transparent, and their specifications can be assessed by standard statistical test procedures. The nature of epidemic trajectories leads us to formulate a class of models in which the logarithm of the growth rate of the cumulative series follows a downward time trend. Allowing this trend to be time-varying introduces flexibility and enables the effects of changes in policy to be tracked and evaluated. The models are able to adapt as the response of the population changes over time. The framework can be extended to modelling the relationship between two or more series. When there is balanced growth, simple regression models can be used to forecast using leading indicators. The models are applicable in a wide range of disciplines.

CO210 Room K2.40 (Hybrid 08) TOPICS IN PARTIAL IDENTIFICATION AND TIME SERIES ECONOMETRICS Chair: Kanchana Nadarajah

C0191: **Bounding program benefits when participation is misreported**
Presenter: **Lina Zhang**, University of Amsterdam, Netherlands
Co-authors: Denni Tommasi

Instrumental variables are commonly used to estimate treatment effects in case of non-compliance. However, the endogenous program participation is often misreported in survey data, and standard techniques are not sufficient to point identify and consistently estimate the effects of interest. We first establish a link between the true and mismeasured effect which is mediated by a parameter of the misclassification probabilities. Second, we provide an instrumental variable method to partially identify the heterogeneous treatment effects when both non-compliance and misreporting of treatment status are present. Third, we formalize a strategy to combine external information about misclassification probabilities of treatment status to tighten the bounds or to obtain a point estimate. Finally, we develop ivbounds, a new Stata package that we use to reassess the benefits of participating in the 401(k) pension plan on savings. Our method has several applications. First, it can be used as the leading strategy in any setting where the practitioner knows that the endogenous binary treatment is not well measured. Second, it can be used as the leading robustness check in case misreporting is only suspected. Third, it can be used to assess the sensitivity of program benefits under different assumptions of the misclassification probabilities.

C0456: **Real-time monitoring of bubbles and crashes**
Presenter: **Emily Whitehouse**, University of Sheffield, United Kingdom
Co-authors: Dave Harvey, Steve Leybourne

Given the financial and economic damage that can be caused by the collapse of an asset price bubble, it is of critical importance to rapidly detect the onset of a crash once a bubble has been identified. We develop a real-time monitoring procedure for detecting a crash episode in a time series. We adopt an autoregressive framework, with the bubble and crash regimes modelled by explosive and stationary dynamics respectively. The first stage is to monitor for the presence of a bubble; conditional on having detected a bubble, we monitor for a crash in real-time as new data emerges. The crash detection procedure employs a statistic based on the different signs of the means of the first differences associated with explosive and stationary regimes, and critical values are obtained using a training period, over which no bubble or crash is assumed to occur. Monte Carlo simulations suggest that the recommended procedure has a well-controlled false positive rate during a bubble regime, while also allowing very rapid detection of a crash when one occurs. Application to the US housing market demonstrates the efficacy of our procedure in rapidly detecting the house price crash of 2006.

C0721: **Bounds on causal direct and indirect average treatment effects in the presence of noncompliance**
Presenter: **Xuan Chen**, Renmin University of China, China

Causal mediation effects are examined in an experimental setting with treatment noncompliance. A binary instrument is used to address noncompliance. Given a particular mechanism variable, we decompose the local average treatment effect (LATE) for individuals who would comply with the assigned treatment (i.e. compliers) into the complier natural indirect (or mechanism) average treatment effect (CMATE), and the complier natural direct (or net) average treatment effect (CNATE). CMATE refers to the part of the LATE that works through the mechanism variable, while CNATE refers to the part that works through all other channels. Without strong assumptions for point identification, we employ the principal stratification approach to derive nonparametric sharp bounds on CMATE and CNATE by imposing weak monotonicity assumptions on mean potential outcomes within or across principal strata, which are subpopulations defined by the joint potential values of the treatment receipt and mechanism under each value of the treatment assignment indicator. We also obtain sharp bounds on the local mechanism and net effects for different strata of compliers. To illustrate the identifying power of our bounds, we analyze the part of the effects of a training program on participants' future earnings and employment that works through the attainment of a vocational degree.

C1057: **Specification tests for GARCH processes with nuisance parameters on the boundary**
Presenter: **Indeewara Perera**, University of Sheffield, United Kingdom

Tests for the correct specification of the conditional variance function in GARCH models are developed when the true parameter may lie on the boundary of the parameter space. The test statistics considered are of Kolmogorov-Smirnov and Cramer-von Mises type, and are based on a certain empirical process marked by centered squared residuals. The limiting distributions of the test statistics are not free from (unknown) nuisance parameters, and hence critical values cannot be tabulated. A novel bootstrap procedure is proposed to implement the tests; it is shown to be asymptotically valid under general conditions, irrespective of the presence of nuisance parameters on the boundary. The proposed bootstrap approach is based on shrinking of the parameter estimates used to generate the bootstrap sample toward the boundary of the parameter space at a proper rate. It is simple to implement and fast in applications, as the associated test statistics have simple closed-form expressions. A simulation study demonstrates that the new tests: (i) have excellent finite sample behaviour in terms of empirical rejection probabilities under the null as well

as under the alternative; (ii) provide a useful complement to existing procedures based on Ljung-Box type approaches. Two data examples are considered to illustrate the tests.

C1158: **Non-parametric variance function estimation in linear regression models with many regressors**
Presenter: **Weilun Zhou**, University of Cambridge, United Kingdom

The nonparametric estimation of unknown variance function in linear regression models that allow for many regressors is considered. When the number of regressors increases at the same rate as the sample size, the conventional nonparametric variance estimator is biased. An orthogonal series expansion is used to approximate the unknown variance function and a leave-one-out slope coefficient estimator is involved to eliminate the bias. The asymptotic properties of the proposed estimator are derived. Simulation evidence consistent with the theoretical results and an empirical illustration is also provided.

CO174 Room K2.41 (Hybrid 09) NEW DEVELOPMENT IN FACTOR MODELS AND THEIR APPLICATIONS Chair: Degui Li

C1317: **Confidence interval construction: A new self-normalization approach based on adjusted range**
Presenter: **Jiajing Sun**, University of Chinese Academy of Sciences, China
Co-authors: Yongmiao Hong, Oliver Linton, Shouyang Wang

A new self-normalization method is proposed to construct confidence intervals for quantities of stationary time series. Unlike the self-normalization approach, which utilizes the variance of a partial as the self-normalizer, we propose the use of its adjusted range instead. Given ranges capacity to deal long-range dependence in highly non-Gaussian time series with large skewness and/or kurtosis, we introduce two range-based autocorrelation tests and study a confidence interval construction for censored dependent data extending previous work. The simulations confirm the validity of our approach.

C0719: **News-implied linkages and local dependency in the equity market**
Presenter: **Shuyi Ge**, University of Nankai, China
Co-authors: Oliver Linton, Shaoran Li

The purpose is to study a heterogeneous coefficient spatial factor model that separately addresses both common factor risks (strong cross-sectional dependence) and local dependency (weak cross-sectional dependence) in the equity returns. From the asset pricing perspective, we derive the theoretical implications of no asymptotic arbitrage for the heterogeneous spatial factor model. In empirical work, it is challenging to measure granular firm-to-firm connectivity for a high-dimensional panel of equity returns. We use extensive business news to construct firms' links via which local shocks transmit, and we use those news-implied linkages as a proxy for the connectivity among firms. Empirically, we document a considerable degree of local dependency among S&P 500 stocks, and the spatial component does a great job in capturing the remaining correlations in the de-factored returns. We find that adding spatial interactions to factor models reduces mispricing and mean-squared errors. We also show that our news-implied linkages provide a comprehensive and integrated proxy for firm-to-firm connectivity, and it out-performs other existing networks in the literature.

C0436: **Estimation of dynamic quantile panel data model with interactive effects**
Presenter: **Chaowen Zheng**, University of York, United Kingdom

The estimation of a dynamic quantile panel data model with unobserved interactive effects is considered. A two-step procedure is proposed. In the first step, we apply the iterative principal component analysis to estimate the unobserved common factors. In the second step, we construct an augmented model using these estimated factors and then run quantile regression to estimate the slope parameters together with the individual effects (factor loadings). To facilitate asymptotic analysis, we smooth the quantile objective function. We show that the proposed two-step estimators are consistent and asymptotically normal-distributed, though subject to asymptotic bias due to the incidental parameter problem. The split-panel jackknife is then employed for correcting the bias. Monte Carlo simulations confirm that our proposed bias-corrected estimator has good finite sample performance.

C0490: **Estimation of common factors for microstructure noise and efficient price in a high-frequency dual-factor model**
Presenter: **Yuning Li**, University of York, China

The Double Principal Component Analysis (DPCA) is developed based on a dual-factor structure for high-frequency intraday returns data contaminated with microstructure noise. The dual-factor structure allows a factor structure for the microstructure noise in addition to the factor structure for efficient log-prices. We construct estimators of factors for both efficient log-prices and microstructure noise and their common components. We provide uniform consistency of these estimators when the number of assets and the sampling frequency go to infinity. In a Monte Carlo exercise, we compare our DPCA method to a PCA-VECM method. Finally, an empirical analysis of intraday returns of S&P 500 Index constituents provides evidence of co-movement of the microstructure noise that distinguishes from latent systematic risk factors.

C1581: **Functional-coefficient quantile regression for panel data with latent group structure**
Presenter: **Degui Li**, University of York, United Kingdom

Estimating functional-coefficient models is considered in panel quantile regression with individual effects, allowing the cross-sectional and temporal dependence for large panel observations. A latent group structure is imposed on the heterogenous quantile regression models so that the number of nonparametric functional coefficients to be estimated can be reduced considerably. With the preliminary local linear quantile estimates of the subject-specific functional coefficients, a classic agglomerative clustering algorithm is used to estimate the unknown group structure and an easy-to-implement ratio criterion is proposed to determine the group number. The estimated group structure and number are shown to be consistent. Furthermore, a post-grouping local linear smoothing method is introduced to estimate the group-specific functional coefficients, and the relevant asymptotic normal distribution theory is derived with a normalisation rate comparable to that in the literature. The developed methodology and theory are applied to identify the latent group structure in linear panel quantile regression (uniformly over quantile levels) and model individual effects.

CC862 Room Virtual R18 CONTRIBUTIONS IN REALIZED VOLATILITY Chair: Hiroyuki Kawakatsu

C1198: **Instrumental realized volatility**
Presenter: **Hiroyuki Kawakatsu**, Dublin City University, Ireland

The use of multiple realized variance measures is considered in observation driven GARCH type and parameter-driven stochastic volatility models. The main idea is to use several noisy realized measurements as instruments of each other to extract signals to help identify the underlying latent variance series. In addition to the commonly used tick, data-based realized measures, information content in the more widely accessible but noisy open-high-low-closed data-based realized measures are considered. The performance of the proposed models is compared using (pseudo) out-of-sample tail risk forecast accuracy.

C1506: **Forecasting realized volatility: A hybrid model integrating BiLSTM with HAR-type models**
Presenter: **Yi Luo**, Lancaster University, United Kingdom
Co-authors: Marwan Izzeldin, Mike Tsionas

In the last few decades, artificial neural networks have been extensively used as a forecasting method for financial time series in both academia and industry. The major advantage of this approach is the possibility to approximate any linear and nonlinear behaviors without knowing the structure

of the data generating process. This makes it suitable for forecasting time series which exhibit long-memory and nonlinear dependencies, like conditional volatility. However, much literature has found that the quality of features fed into the neural networks is crucial to the success of the performance of such models. A hybrid methodology is proposed that combines both heterogeneous autoregressive (HAR)-type models and deep feedforward neural network (DFN) model as well as bidirectional long short-term memory (BiLSTM) model in predicting realized volatility. The results show that BiLSTM-based hybrid model outperforms all other models in out-of-sample forecasting. Additionally, the performance of both DFN-based and BiLSTM-based hybrid model beat their single model counterparts, indicating HAR-type components can be served as effective features in DFN and BiLSTM structure.

C1491: **Volatility spillover among Japanese sectors in response to COVID-19**
Presenter: **Hideto Shigemoto**, Kwansei Gakuin University, Japan
Co-authors: Takayuki Morimoto
The aim is to clarify how risks spread across economic sectors and indicate the sectors that are most affected compared to the others in order to help investors with asset allocation and support them in risk management. Although the Japanese stock market is one of the relatively large stock markets in the world, there have been no studies on volatility spillovers among its sectors. The volatility spillovers among 17 sectors classified in the Tokyo Stock Exchange are examined by using the forecast error variance decomposition of the vector autoregressive model. The results show that the pattern of volatility spillovers across sectors in the Japanese stock market differs between the pre-COVID-19 and the during the COVID-19 period. While energy resources and bank sectors are risk receivers in the pre-COVID-19 period, these sectors are risk transmitters during the COVID-19 period. We also find that volatility spillovers in the Japanese stock market are mainly driven by negative realized semi-variance. These results are useful for asset allocation and risk management.

C1123: **Predictive power of the variance premium**
Presenter: **Yuze Liu**, Fernuniversitat in Hagen, Germany
The predictive power of the variance premium (VP) on the equity premium is tested in different aspects for the S&P 500 index. The VP is defined as the difference between implied volatility and the physical expectation of conditional variance. We consider various versions of the VP: VIX, Structural VIX and Corridor VIX are employed as implied volatility measurement. 5-min realized volatility (RV) forecasts in an extended HAR framework including jumps and daily quadratic variation and the recently MF2-GARCH model featuring leverage effect and the lower frequency of conditional variance are considered as the physical measures. First, the forecasting performance of different conditional variance measures is compared to select the best physical measurement. Second, an extensive comparison of the predictive power is performed by using different VP measures. The performance is evaluated by in-sample and out-of-sample predictive power and predictive regression tests. Third, we discuss the difference in the predictive power between VP, implied volatility and conditional volatility. The results are discussed in light of the usefulness of newly proposed VP measures in comparison to implied or conditional volatility, respectively.

C1473: **A volatility spillover analysis with realized semi(co)variances in Australian electricity markets**
Presenter: **Ainhoa Zarraga**, University of the Basque Country, UPV/EHU, Spain
Co-authors: Aitor Ciarreta, Evelyn Lizeth Chanatasig
Volatility spillovers are a characteristic of interconnected electricity markets. We use high-frequency prices to analyze the transmission of volatility across five Australian regional electricity markets. We decompose the realized covariance matrix based on the sign of the underlying returns and propose three models. The first includes only variances, the second adds covariances and the third includes semi(co)variances obtained from the decomposition of the realized covariance matrix. We carry out the analysis for both a static and a dynamic framework and relate the behavior of spillovers to major events and policies affecting the market. Results show a high level of integration across markets and highlight the importance of the role of semi(co)variances in detecting asymmetric spillovers.

CC861 Room Virtual R31 CONTRIBUTIONS IN BAYESIAN ECONOMETRICS Chair: Lukasz Kwiatkowski

C1557: **Bayesian assessment of identifying restrictions for heteroskedastic structural VARs**
Presenter: **Tomasz Wozniak**, University of Melbourne, Australia
A flexible Bayesian structural vector autoregressive model is introduced identified through heteroskedasticity, encompassing a range of volatility processes and allowing for additional identifying restrictions. Consequently, it enables comparisons across structural models with alternative sets of restrictions that just identify homoskedastic specifications. The structural model exhibits a new standardization that sets the sum of each structural shock variances to one. This solution facilitates the development of a complete toolset for Bayesian inference, including a reference prior, an efficient estimation algorithm, and an unbiased marginal data density estimator for locally identified models. Applying this apparatus to three U.S. monetary policy models, we document the empirical outperformance of models making use of two policy variables over those with a single one.

C1663: **Estimating Bayesian models using simulated data meta-learning**
Presenter: **Sergei Seleznev**, Bank of Russia, Russia
Co-authors: Ramis Khabibullin
A simple algorithm is presented for the estimation of Bayesian models that is based on the principles of meta-learning literature. The algorithm consists of two main steps: artificial data generation and fitting a neural network to the variables of interest. In the first step, an artificial dataset is created as a set of samples from the joint distribution of parameters (generated from prior) and data (generated from the data generating process). In the second step, the neural network is trained in a supervised manner to predict variables of interest (such as parameters and/or hidden variables) on the previously generated dataset. It is shown that the algorithm converges to the posterior mean or any other characteristic of posterior distribution depending on the loss function used in the training of the neural network. The main advantage of the proposed method is that trained once it can be used for any dataset without additional training, so the inference of the Bayesian model becomes almost instantaneous. Two examples (stochastic volatility model and new seasonal adjustment procedure) illustrate algorithm properties.

C1743: **Media bias and polarization using a Markov-switching latent space network model**
Presenter: **Antonio Peruzzi**, Ca' Foscari University of Venice, Italy
The news consumption landscape has drastically changed in the last decades. Web 2.0 and social media are re-shaping the way in which news pieces are consumed and produced. Some old questions renew in such a scenario. One of these is whether and to which extend news outlets bias information. We propose a new dynamic latent-space model (LS) for news outlets in which we exploit both time-varying online duplication-network data as well as textual contents from published articles to measure media bias over time. Within our model, the latent-space positions of news outlets have a proper interpretation respectively in terms of media slant and online engagement. The aim is twofold: making advancements both concerning the analysis of the timely evolution of audience duplication networks and concerning the determination of media slant and polarization by exploiting both textual and network-structure information. The developed model is applied to a Facebook dataset consisting of the information provided by Italian news outlets in the years 2015 and 2016. Eventually, the latent positions of the news outlets over time is analyzed and discussed.

C1316: **Comparison of Bayesian models in the context of recursive density predictions on the example of VEC-SV-GARCH models**
Presenter: **Lukasz Kwiatkowski**, Cracow University of Economics, Poland
Co-authors: Anna Pajor, Jacek Osiewalski, Justyna Wroblewska
The focus is on a formal Bayesian method of recursive multi-step-ahead density prediction and its ex-post evaluation. We propose a new decompo-

sition of the so-called predictive Bayes factor of order (k,s) into the product of partial Bayes factors. The first factor in the decomposition (called the predictive Bayes factor of order k) is related to the relative k-period-ahead forecasts ability of models, and the second factor is connected with the recursive updates of posterior odds ratios based on updated data sets. To illustrate the usefulness of the measures proposed, we apply the new decomposed predictive Bayes factors to compare the forecasting ability of models when the true data generating process (DGP) is known. The simulation results suggest that the predictive Bayes factor of order (k,s) introduced here and accounting for the updating effect allows pinpointing the model based on the true DGP. Next, we investigate the predictive ability of different vector error correction models with heteroscedasticity (stochastic volatility and generalized autoregressive conditional heteroskedasticity structures) for sets of the US and Polish macroeconomic variables: unemployment, inflation and interest rates. The results show that the forecasting ability of the models depends on the forecast horizon as well as on taking into account the updating effect.

CC863 Room Virtual R32 CONTRIBUTIONS IN MACRO AND FINANCE I Chair: Alessia Paccagnini

C1355: **The ECB's policy, the recovery fund and the importance of trust: The case of Greece**
Presenter: **Vasiliki-Eirini Dimakopoulou**, Athens University of Economics and Business, Greece
Co-authors: Apostolis Apostolis Philippopoulos, George Economides

The purpose is, using a microfounded macroeconomic model that embeds the key features of the Greek economy, to study the efficacy of the various policy measures taken, at a national and EU level, to cushion the economic effects of the pandemic shock. We attempt to give quantitative answers to questions like: What are the effects of these policies and, especially, what are the implications of the fiscal transfers and grants from the Recovery Fund and the quantitative policies of the ECB, like the PEPP, for the Greek economy? Do they help the real economy and, if yes, by how much? What would have happened had these measures not been taken? How costly will be the re-emergence of the fear of debt default and risk premia? V. Dimakopoulou is grateful to the State Scholarships Foundation (IKY) for financial support. This research is co-financed by Greece and the European Union (European Social Fund- ESF) through the Operational Programme "Human Resources Development, Education and Lifelong Learning" in the context of the project Reinforcement of Postdoctoral Researchers - 2nd Cycle (MIS-5033021), implemented by the State Scholarships Foundation (IKY).

C0264: **Macroeconomic effects of collateral requirements and financial shocks**
Presenter: **Aicha Kharazi**, Free University of Bozen-Bolzano, Italy

A standard model of the business cycle is introduced to study the dynamic effects of collateral requirements. The model is estimated using Bayesian methods and can be employed to ask whether an exogenous change in a firm's ability to borrow can amplify macroeconomic fluctuations via the bank lending channel and to measure the role of collateral. Regarding the results from the estimated model, first, we find that the effect of an increase in collateral requirements is highly significant, while the effect of investment-specific technical change is remarkably modest compared with the other financial shocks. When financial data, such as external financing, business net worth, and capital price, are incorporated into the model, productivity shocks diminish in importance. Second, we find that the contribution of financial shocks is much marked during the financial crisis and substantially shapes macroeconomic fluctuations.

C1208: **Time to see a doctor: Expenditure at retirement in Japan**
Presenter: **Kento Tango**, Yokohama City University, Japan
Co-authors: Yoshiyuki Nakazono

Using household panel data, it is tested whether there exists a decline in consumption at retirement. We find stark evidence of retirement consumption; there is an immediate decline in expenditure of 2.4% even at expected retirement. The negative effect of retirement on expenditure is persistent, and it lasts for at least two years. However, there is no dip in the consumption of higher-educated households, as is the case with lower-educated households. Further, the decline in the consumption of healthcare products such as drugs is severe. Indeed, lower-educated households decrease expenditure on drugs by 25% at most. An additional survey for healthcare use reveals that frequent visits to the doctor explain the decline in expenditure on over-the-counter (OTC) drugs. The results suggest that the reduced opportunity cost of time to see a doctor induces households at retirement to visit a doctor more often than before, and obtain prescribed drugs at more affordable prices than OTC drugs, possibly owing to the universal health insurance system.

C1171: **A flexible approach towards time-varying parameter VAR models**
Presenter: **Boris Blagov**, RWI - Leibniz Institute for Economic Research, Germany

A flexible Bayesian VAR framework is proposed which incorporates regime-switching along with time-varying parameters and develops a new estimation method for Markov-switching models that permits inference even when a regime is not visited. The usefulness of this approach is evaluated in a Monte Carlo exercise with artificial data.

C1356: **Structural estimation combining micro and macro data**
Presenter: **Luca Neri**, Aarhus University, Denmark

A novel approach is introduced for estimating heterogeneous-agent macroeconomic models adding information from micro data. The methodology covers both panels and repeated cross sections, with applications to a wide class of dynamic structural models used in macroeconomics. The routine involves the estimation of dynamic moments over subgroups of the cross-sectional dimension of agents. Additionally it proposes a method for comparison of alternative choices of sub-population moments for the estimation of deep parameters of the economy. Micro moments differ from each other in the informative content that they carry for point estimation of the structural parameters. For instance, variability of moments over the cross-sectional distribution of households' wealth contains relevant information for the correct estimation of the subjective discount rate. However, data from the cross section are not relevant for the identification of a technology shock.

CC876 Room Virtual R34 CONTRIBUTIONS IN FINANCIAL ECONOMETRICS III Chair: Shan Lu

C0357: **VIX derivatives valuation: The effects of jump contagion**
Presenter: **Shan Lu**, University of Kent, United Kingdom
Co-authors: Euan Phimister

The purpose is to analyze whether allowing jump contagion improves the performance of VIX option pricing models in terms of pricing and hedging. Based on an affine jump-diffusion with double jumps, we consider a general jump structure that subsumes both Poisson and Hawkes jumps, as well as models with diffusion-driven jumps, two vol-of-vol factors, and a model that allows for contagions between negative and positive jumps. We derive the semi-closed form option pricing formulas and local minimum-variance hedge ratios for the models. We show that, overall, allowing for jump contagion improves the model performance.

C0324: **A multicountry model of the term structures of interest rates with a GVAR**
Presenter: **Rubens Moura**, Universita catholique de Louvain, Belgium
Co-authors: Candelon Bertrand

Global interdependencies have caused affine term structure models (ATSMs) to adopt a multicountry dimension. Nevertheless, recent referenced ATSMs face issues of tractability as the model dimension becomes larger. To close this gap, an ATSM is proposed in which the risk factor dynamics follow a global vector autoregressive (GVAR). ATSM-GVAR renders a parsimonious yield curve parametrization, which allows for

a fast estimation process, enables meaningful statistical inference of economic relationships, and produces accurate bond yields out-of-sample forecasting. To empirically illustrate our novel ATSM, we build a markedly integrated economic system composed of three Latin American economies and China. We find that, consequent to its prominent role in the worldwide economy, China's economic stances have non-negligible impacts on Latin American yield curve dynamics.

C0595: **Global and regional factors and the term structure of interest rates: Some evidence from Asian countries**

Presenter: **Takeshi Kobayashi**, NUCB Business School, Japan

Asian bond markets have rapidly progressed during the last two decades after the 1997-98 financial crisis in this region. Our approach has extended a previous global factor model to examine advanced countries and Asian countries simultaneously during 2003-2020. We use a two-step state-space model to extract the global, regional and local factors. We show that the global level factor explains on average 50% of the variations of level factors of Asian countries. The results also demonstrate that while there is a significant regional impact on Asian countries, the degree of regional influence differs across these countries. Moreover, we investigate relationships between the macro-economic variables and global and regional factors. It contributes to the literature by examining the degree of integration between Asian and international bond markets. These findings would help bond portfolio managers to be concerned with their country allocation decision.

C0911: **Immunization with term structure dynamics**

Presenter: **Jorge Wolfgang Hansen**, Aarhus University and CREATES and the Danish Finance Institute, Denmark

Co-authors: Daniel Borup, Bent Jesper Christensen

It is shown that improved hedging performance of bond portfolios can be obtained by exploiting no-arbitrage and consistency restrictions on the interest rate dynamics rather than using purely cross-sectional approaches based on factor analysis or a parametrized yield curve shape. We also introduce a new term structure model involving three stochastically varying factors corresponding to level, slope, and curvature, which leads to a further improvement in hedging performance. A restricted version of this model belongs to the three-factor affine class. The evidence shows the importance of immunizing the factor contribution to hedging error variance contrary to balancing factor versus idiosyncratic contributions.

C0583: **New stylized facts of financial exuberance**

Presenter: **Marco Kerkemeier**, University of Hagen, Germany

Co-authors: Christoph Wegener, Robinson Kruse-Becher

Financial asset prices exhibit temporary mild explosive behaviour. We investigate the behaviour of daily (real) financial returns during exuberance by using a broad data set consisting of equity indices, precious metals, oil, real estate and cryptocurrencies. Adding to common knowledge about stylized facts during normal market phases, we explicitly study the typical behaviour of returns during these explosive periods. Explosive phases are identified and date-stamped by using the popular Dickey-Fuller based procedures. We distinguish rising prices from collapsing phases. Returns of both subperiods are studied for recurring characteristics. Results for explosive phases are benchmarked against nonexplosive periods. We study features of explosive phases, i.e. duration, magnitude of explosiveness and types of collapsing behaviour. We apply a battery of econometric procedures to establish potential stylized facts (e.g., autocorrelation, distributional characteristics, volatility persistence). Our findings are applied to risk management, in particular to Value at Risk and Expected Shortfall estimation and forecasting. Especially collapsing phases are relevant in this context. We investigate to what extent improvements are achievable by accounting for potentially different characteristics during explosive phases. Finally, in a Monte Carlo simulation study, we reinvestigate the empirical power of popular bubble tests under realistic circumstances matching our new findings.

Sunday 19.12.2021 14:00 - 15:40 Parallel Session I – CFE-CMStatistics

EO750 Room K0.16 (Hybrid 02) BFF: TOPICS IN FOUNDATIONS OF INFERENCE Chair: Jan Hannig

E0435: **Confidence in Bayesian and fiducial inference**
Presenter: **Gunnar Taraldsen**, NTNU Norway, Norway

Fiducial inference has not been widely accepted. Different versions are presented in textbooks and most often with critical remarks. Fiducial inference was, in fact, declared dead at the end of the last century. There is little confidence in fiducial inference. In contrast, others suggested that maybe Fisher's biggest blunder will become a big hit in the 21st century. One reason for this optimism is the version of fiducial inference based on a data-generating model. This is perfectly adapted to the available computational power in the 21st century. We have much confidence in fiducial inference. We present results that ensure that a fiducial distribution is a confidence distribution. Similarly, conditions are given that ensure that a fiducial is a Bayes posterior. Our main example is the correlation in a binormal distribution as used initially in 1930 when Fisher introduced the fiducial argument. An exact formula for the confidence density of the correlation has recently been derived, and this allows illustrating the theory more directly than before.

E0464: **BFF: Bayesian, fiducial, and frequentist analysis of cognitive engagement among cognitively impaired older adults**
Presenter: **Shevaun Neupert**, North Carolina State University, United States
Co-authors: Claire Growney, Xianghe Zhu, Julia Sorensen, Emily Smith, Jan Hannig

Engagement in cognitively demanding activities is beneficial in preserving cognitive health. The goal was to demonstrate the utility of frequentist, Bayesian, and fiducial statistical methods for evaluating the robustness of effects in identifying factors that contribute to cognitive engagement for older adults experiencing cognitive decline. We collected a total of 504 observations across two longitudinal waves of data from 28 cognitively impaired older adults. Participants systolic blood pressure response, an index of cognitive engagement, was continuously sampled during cognitive testing. Participants reported on physical and mental health challenges and provided hair samples to assess chronic stress at each wave. Using the three statistical paradigms, we compared results from models testing levels and longitudinal changes in health and stress predicting changes in cognitive engagement. Findings were mostly consistent across the three paradigms, providing additional confidence in determining effects. We emphasize the utility of the Bayesian and fiducial paradigms for use with relatively small sample sizes because they are not based on asymptotic distributions. In particular, a fiducial paradigm is a useful tool because it provides more information than p values without the need to specify prior distributions, which may unduly influence the results based on a small sample.

E1135: **A Gibbs sampler for a class of random convex polytopes**
Presenter: **Ruobin Gong**, Rutgers University, United States

A Gibbs sampler is presented for the Dempster-Shafer (DS) approach to statistical inference for Categorical distributions. The DS framework extends the Bayesian approach, allows in particular the use of partial prior information, and yields three-valued uncertainty assessments representing probabilities "for", "against", and "don't know" about formal assertions of interest. The proposed algorithm targets the distribution of a class of random convex polytopes which encapsulate the DS inference. The sampler relies on an equivalence between the iterative constraints of the vertex configuration and the non-negativity of cycles in a fully connected directed graph.

E0432: **Calibration of likelihood ratios systems in forensic science**
Presenter: **Jan Hannig**, University of North Carolina at Chapel Hill, United States
Co-authors: Hari Iyer

Many computer programs and software systems used in the interpretation of forensic evidence have as their output Bayes factors, also commonly referred to as likelihood ratios. For example, it is not unusual to see it reported that the DNA recovered at the crime scene is a million times more likely under the assumption that the defendant is a contributor to the crime stain than under the assumption that the defendant is not a contributor. We summarize existing approaches for examining the validity of likelihood ratio systems. We will see that what is used in the current practice has significant drawbacks related to uncertainty quantification. We then discuss a new statistical methodology based on generalized fiducial inference for empirically examining the validity of such likelihood ratio assessments. Using data from a number of sources, such as glass, paint and DNA evidence, we illustrate our approach by examining LR values calculated using standard approaches in the forensic literature. We also use the new tool to show limitations of a common method of post-hoc re-calibrating of outputs

EO306 Room K0.18 (Hybrid 03) VARIABLE SELECTION IN CAUSAL INFERENCE Chair: Richard Guo

E1231: **Debiased IPW for estimation of average treatment effects with high-dimensional confounders**
Presenter: **Yuhao Wang**, Tsinghua University, China
Co-authors: Rajen D Shah

Estimation of average treatment effects given observational data with high-dimensional pretreatment variables is considered. Existing methods for this problem typically assume some form of sparsity for the regression functions. We introduce a debiased inverse propensity score weighting (DIPW) scheme for average treatment effect estimation that delivers $\sqrt{n}$-consistent estimates of the average treatment effect when the propensity score follows a sparse logistic regression model; the regression functions are permitted to be arbitrarily complex. Our theoretical results quantify the price to pay for permitting the regression functions to be unestimable, which shows up as an inflation of the variance of the estimator compared to the semiparametric efficient variance by at most $O(1)$ under mild conditions. Given the lack of assumptions on the regression functions, averages of transformed responses under each treatment may also be estimated at the $\sqrt{n}$ rate, and so for example, the variances of the potential outcomes may be estimated. We show how confidence intervals centred on our estimates may be constructed, and also discuss an extension of the method to estimating projections of the heterogeneous treatment effect function.

E1297: **Causal ball screening: Outcome model-free variable selection for causal inference**
Presenter: **Dingke Tang**, University of Toronto, Canada
Co-authors: Dehan Kong, Wenliang Pan, Linbo Wang

Causal inference has been increasingly reliant on observational studies with rich covariate information. To build tractable causal procedures, including the doubly robust estimation, it is imperative to first extract important features from high dimensional data. Unlike the familiar task of variable selection for prediction modelling, our feature selection procedure aims to control for confounding while maintaining efficiency in the resulting causal effect estimate. Previous empirical and theoretical studies imply that one should aim to include all predictors of the outcome, rather than the treatment, in the propensity score model. We formalize this intuition through rigorous proofs and propose the causal ball screening for selecting these variables from modern ultra-high dimensional data sets. A distinctive feature of our proposal is that our procedure is more efficient than existing methods while our procedure keeps the desirable double robustness property. Our theoretical analyses show that the proposed procedure enjoys a number of favorable properties, including model selection consistency, normality and efficiency. Synthetic and real data analyses show that our proposal performs favorably with existing methods in a range of realistic settings.

E1396: **Post-machine learning causal inference: Uniformly valid inference and sensitivity analysis**
Presenter: **Xavier de Luna**, Umea University, Sweden
Co-authors: Niloofar Moosavi, Tetiana Gorbach, Jenny Haggstrom

The recent literature on uniform valid causal inference is reviewed, and sensitivity analysis to unobserved confounders is discussed in this context. When inference is aimed at a low dimensional causal parameter, double robust estimation strategies combined with machine learning algorithms to fit high-dimensional nuisance models are becoming increasingly popular. This popularity is due to key theoretical results obtained in recent years that guarantee the uniform validity of the inference post-model selection (or post-machine learning). We discuss the costs and benefits of using uniformly valid causal inference. Another major threat to the validity of causal inference is the potential existence of unobserved confounders. We, therefore, present new results for dealing with uncertainty due to unobserved confounding in the context of uniformly valid causal inference.

E1422: **Variable elimination and graph reduction: Towards an efficient g-formula**
Presenter: **Richard Guo**, University of Cambridge, United Kingdom
Co-authors: Emilija Perkovic, Andrea Rotnitzky

Consider a study where the causal structure is known and described by a directed acyclic graph (DAG). A causal quantity of interest, say a counterfactual mean, can often be expressed as a functional of the observed distribution given by the g-formula (also known as the "truncated factorization"). The g-formula, which can be written down from the graph, usually takes the form of an integral involving conditional expectations of the variables in the graph. Naturally, to estimate the causal quantity efficiently, one can use a plugin estimator of the g-formula, where every conditional expectation is replaced by its MLE. However, we find that asymptotically not every variable appearing in the g-formula carries information for estimation. In fact, the causal quantity can often be estimated with an "efficient" g-formula that drops the redundant variables such that the cost of measuring these variables can be saved. We present a graphical procedure towards this goal. First, we identify a set of graphical conditions that are necessary and sufficient for eliminating redundant variables. Second, we construct a reduced DAG on the non-redundant variables, from which the "efficient" g-formula can be derived. The reduced DAG is transformed from the original DAG through a set of moves, traversing both within and between Markov equivalence classes, which nonetheless leave the semiparametric efficiency bound of the estimation problem invariant.

EO208 Room K0.19 (Hybrid 04) RECENT ADVANCES IN HIGH-DIMENSIONAL DATA ANALYSIS Chair: Zijian Guo

E0679: **Surrogate assisted semi-supervised inference for high dimensional risk prediction**
Presenter: **Jue Hou**, Harvard T.H. Chan School of Public Health, United States
Co-authors: Zijian Guo, Tianxi Cai

Risk modeling with EHR data is challenging due to a lack of direct observations on the disease outcome, and the high dimensionality of the candidate predictors. We develop a surrogate assisted semi-supervised-learning (SAS) approach to risk modeling with high dimensional predictors, leveraging a large unlabeled data on candidate predictors and surrogates of the outcome, as well as a small labeled data with annotated outcomes. The SAS procedure borrows information from surrogates along with candidate predictors to impute the unobserved outcomes via a sparse working imputation model with moment conditions to achieve robustness against miss-specification in the imputation model and a one-step bias correction to enable interval estimation for the predicted risk. We demonstrate that the SAS procedure provides valid inference for the predicted risk derived from a high dimensional working model, even when the underlying risk prediction model is dense and the risk model is miss-specified. We present an extensive simulation study to demonstrate the superiority of our SSL approach compared to existing supervised methods. We apply the method to derive genetic risk prediction of type-2 diabetes mellitus using an EHR biobank cohort.

E0541: **Optimal and safe estimation for high-dimensional semi-supervised learning**
Presenter: **Yang Ning**, Cornell University, United States

The estimation problem in high-dimensional semi-supervised learning is considered. The goal is to investigate when and how the unlabeled data can be exploited to improve the estimation of the regression parameters of a linear model in light of the fact that such linear models may be misspecified in data analysis. We first establish the minimax lower bound for parameter estimation in the semi-supervised setting. We show that the supervised estimators using the labeled data only cannot attain this lower bound. When the conditional mean function is correctly specified, we propose an optimal semi-supervised estimator which attains the lower bound and therefore improves the rate of the supervised estimators. To alleviate the strong requirement for this optimal estimator, we further propose a safe semi-supervised estimator. We view it safe, because this estimator remains minimax optimal when the conditional mean function is correctly specified, and is always at least as good as the supervised estimators. Furthermore, we extend our idea to aggregate multiple semi-supervised estimators caused by different misspecifications of the conditional mean function. Extensive numerical simulations and a real data analysis are conducted to illustrate our theoretical results.

E1412: **The projected covariance measure for model-free variable significance testing**
Presenter: **Rajen D Shah**, University of Cambridge, United Kingdom
Co-authors: Ilmun Kim, Anton Rask Lundborg, Richard Samworth

Testing the significance of a variable X for predicting a response Y given additional covariates Z, is a ubiquitous task in statistics. One approach is to specify a generalised linear model, and then test whether the regression coefficient for X is non-zero. However, when the model is misspecified, as will invariably be the case, the test may have poor power, for example when X is involved in complex interactions, or lead to many false rejections. We study the problem of testing the model-free null that the conditional mean of Y given X and Z does not depend on X. We propose a simple and general framework that can leverage flexible machine learning methods such as random forests or neural nets to yield both robust error control and high power. The procedure involves performing 4 regressions, two to construct a particular projection of Y on X and Z using one half of the data, and the remaining two to estimate the expected conditional covariance between this projection and Y on the remaining half of the data. By using appropriate regression methods, we show that settings, where Z or X are high-dimensional, can be tackled when there is an underlying sparse model. In the case where X and Z are of moderate dimension, we show that a version of our procedure using spline regression achieves (up to a log factor) what we prove is the minimax optimal rate for nonparametric testing.

E1721: **Variational inference and robustness**
Presenter: **Cynthia Rush**, Columbia University, United States

Variational inference (VI) is a machine learning technique that approximates difficult-to-compute probability densities by using optimization. While VI has been used in numerous applications, it is particularly useful in Bayesian statistics where one wishes to perform statistical inference about unknown parameters through calculations on a posterior density. We will discuss some new ideas about VI and robustness to model misspecification. In particular, we will study alpha-posteriors, which distort standard posterior inference by downweighting the likelihood, and their variational approximations. We will see that such distortions, if tuned appropriately, can outperform standard posterior inference when there is potential parametric model misspecification.

EO074 Room K0.20 (Hybrid 05) STATISTICAL METHODS FOR HIGH DIMENSIONAL NEUROIMAGING DATA I Chair: John Kornak

E0779: **A time-varying AR, bivariate DLM of functional near-infrared spectroscopy data**
Presenter: **Timothy Johnson**, University of Michigan, United States

Functional near-infrared spectroscopy (fNIRS) is a relatively new neuroimaging technique. It is a low cost, portable, and non-invasive method to measure brain activity via the blood oxygen level-dependent signal. Similar to fMRI, it measures changes in the level of blood oxygen in the brain. Its time resolution is much finer than fMRI, however, its spatial resolution is much 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. Furthermore, fNIRS measures the concentration of both oxygenated and deoxygenated hemoglobin, both of which may be of scientific interest. We propose a fully Bayesian time-varying autoregressive model to analyze fNIRS data within the multivariate DLM framework. The hemodynamic response function is modeled with the canonical HRF and the low-frequency drift with a variable B-spline model (both locations and number of knots are allowed to vary). Both the model error and the auto-regressive processes vary with time. Via simulation studies, we show that this model naturally handles motion artifacts and gives good statistical properties. The model is then applied to an fNIRS data set.

E0817: **Efficient Bayesian estimation of brain activation with cortical surface and subcortical data using EM**

Presenter: **Daniel Spencer**, Indiana University, United States

Co-authors: Amanda Mejia, David Bolin, Mary Beth Nebel

Analysis of brain imaging scans is critical to understanding the way the human brain functions. In particular, functional magnetic resonance imaging (fMRI) scans give detailed data on a living subject at relatively high spatial and temporal resolutions. Due to the high cost involved in the collection of these scans, robust methods of analysis are of critical importance in order to produce meaningful inference. Bayesian methods, in particular, allow for the inclusion of expected behavior from a prior study into an analysis, increasing the power of the results while circumventing problems that arise in classical analyses, including the effects of smoothing results and sensitivity to multiple comparison testing corrections. Recent work developed a surface-based spatial Bayesian general linear model (GLM) for cortical surface fMRI (cs-fMRI) data using stochastic partial differential equation (SPDE) priors which rely on the computational efficiencies of the integrated nested Laplace approximation (INLA) to perform powerful analyses. We develop an exact Bayesian analysis method for the GLM, employing an expectation-maximization (EM) algorithm to find estimates of effects of task-based regressors on cs-fMRI and subcortical fMRI data. Our proposed method is compared to the INLA implementation of the Bayesian GLM, as well as the classical GLM on simulated data. An analysis of data from the Human Connectome Project is also shown.

E0849: **Adaptive regularization with applications to brain imaging**

Presenter: **Jaroslaw Harezlak**, Indiana University School of Public Health-Bloomington, United States

The problem of adaptive incorporation of multi-modal brain imagining data sources in multiple linear regression settings is addressed. In the presented example, we model scalar outcome dependence on the brain cortical properties, e.g. cortical thickness and cortical area. We utilize both connectivity and spatial proximity information to build adaptive penalty terms in the regularized regression problem. The general idea of incorporating external information in the regularization approach via linear mixed model representation has been recently established in our prior proposal named ridgefield Partially Empirical Eigenvectors for Regression (riPEER). We incorporate multiple sources of information, including structural and functional connectivity network structure as well as the spatial distance between the cortical regions to estimate the regression parameters with multiple penalty terms via a riPEER extension called AIMER (Adaptive Information Merging Estimator for Regression). We present a simulation study testing various realistic scenarios and apply AIMER to data arising from the Human Connectome Project (HCP) study.

E0224: **Permutation testing of network enrichment in neuroimaging**

Presenter: **Sarah Weinstein**, University of Pennsylvania, United States

Co-authors: Aaron Alexander-Bloch, Simon Vandekar, Erica Baller, Ruben Gur, Raquel Gur, Armin Raznahan, Azeez Adebimpe, Theodore Satterthwaite, Russell Shinohara

In studies of neurodevelopment, anatomical and functional brain parcellations are often used to analyze patterns of association between a phenotype (e.g., age) and an imaging feature (e.g., cortical thickness) and determine whether those associations are especially strong or "enriched" in those subregions. However, existing methods in this area may not be reliable, as they do not account for the spatial structure of the data, leading to inflated type I errors. We propose Network Enrichment Analysis Testing (NEAT), which adapts Gene Set Enrichment Analysis (GSEA), widely used in genomics research, to statistically test whether associations between high-dimensional imaging features and non-imaging phenotypes are enriched within functional networks or other parcellations of the brain. NEAT incorporates random permutations of subject-level data and augments imaging measurements with spatial smoothing to enhance statistical power in the assessment of network-specific enrichment, while preserving conservative type I error rates. We illustrate the properties of NEAT in simulated and real neuroimaging data from the Philadelphia Neurodevelopmental Cohort.

EO234 Room Virtual R18 EXPERIMENTAL DESIGN IDEAS FOR MACHINE LEARNING Chair: HaiYing Wang

E1216: **Subdata selection algorithm for linear model selection**

Presenter: **Jun Yu**, Beijing Institute of Technology, China

A statistical method is likely to be sub-optimal if the assumed model does not reflect the structure of the data at hand. For this reason, it is important to perform model selection before statistical analysis. However, selecting an appropriate model from a large candidate pool is usually computationally infeasible when faced with a massive data set, and little work has been done to study data selection for model selection. We propose a subdata selection method based on leverage scores which enables us to conduct the selection task on a small subdata set. The method not only improves the probability of selecting the best model but also enhances the estimation efficiency. Several examples are presented to illustrate the proposed method.

E1264: **Supervised compression of big data**

Presenter: **Simon Mak**, Duke University, United States

The phenomenon of big data has become ubiquitous in nearly all disciplines, from science to engineering. A key challenge is the use of such data for fitting statistical and machine learning models, which can incur high computational and storage costs. One solution is to perform model fitting on a carefully selected subset of the data. Various data reduction methods have been proposed in the literature, ranging from random subsampling to optimal experimental design-based methods. However, when the goal is to learn the underlying input-output relationship, such reduction methods may not be ideal, since it does not make use of the information contained in the output. To this end, we propose a supervised data compression method called supercompress, which integrates output information by sampling data from regions most important for modeling the desired input-output relationship. An advantage of supercompress is that its nonparametric compression method does not rely on parametric modeling assumptions between inputs and output. As a result, the proposed method is robust to a wide range of modeling choices. We demonstrate the usefulness of supercompress over existing data reduction methods, in both simulations and a taxicab predictive modeling application.

E1451: **Orthogonal subsampling for big data linear regression**

Presenter: **Lin Wang**, George Washington University, United States

The dramatic growth of big datasets presents a new challenge to data storage and analysis. Data reduction, or subsampling, that extracts useful information from datasets is a crucial step in big data analysis. We propose an orthogonal subsampling (OSS) approach for big data with a focus on linear regression models. The approach is inspired by the fact that an orthogonal array of two levels provides the best experimental design for linear regression models in the sense that it minimizes the average variance of the estimated parameters and provides the best predictions. The merits of OSS are three-fold: (i) it is easy to implement and fast; (ii) it is suitable for distributed parallel computing and ensures the subsamples selected in different batches have no common data points; and (iii) it outperforms existing methods in minimizing the mean squared errors of the estimated parameters and maximizing the efficiencies of the selected subsamples. Theoretical results and extensive numerical results show that the OSS approach is superior to existing subsampling approaches. It is also more robust to the presence of interactions among covariates and, when they do exist, OSS provides more precise estimates of the interaction effects than existing methods. The advantages of OSS are also illustrated

through analysis of real data.

EO734 Room Virtual R20 RECENT ADVANCES IN GRAPHICAL MODELS AND DIMENSION REDUCTION Chair: Eftychia Solea

E0312: **On sufficient graphical models**
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, our graphical model is based on conditional independence given a set of sufficient predictors with a substantially reduced dimension. In this way, we avoid the curse of dimensionality that comes with a high-dimensional kernel. We develop the population-level properties, convergence rate, and variable selection consistency of our estimate. By simulation comparisons and an analysis of the DREAM 4 Challenge data set, we demonstrate that our method outperforms the existing methods when the Gaussian or copula Gaussian assumptions are violated. Its performance remains excellent in the high-dimensional setting.

E0810: **On skewed Gaussian graphical models**
Presenter: **Tianhong Sheng**, The Pennsylvania State University, United States
Co-authors: Bing Li, Eftychia Solea

A skewed Gaussian graphical model is introduced as an extension to the Gaussian graphical model. One of the appealing properties of the Gaussian distribution is that conditional independence can be fully characterized by the sparseness in the precision matrix. The skewed Gaussian distribution adds a shape parameter to the Gaussian distribution to take into account possible skewness in the data; thus it is more flexible than the Gaussian model. Nevertheless, the appealing property of the Gaussian distribution is retained to a large degree: the conditional independence is still characterized by the sparseness in the parameters, which now include a shape parameter in addition to the precision matrix. As a result, the skewed Gaussian graphical model can be efficiently estimated through a penalized likelihood method just as the Gaussian graphical model. We develop an algorithm to maximize the penalized likelihood based on the alternating direction method of multipliers, and establish the asymptotic normality and variable selection consistency for the new estimator. Through simulations, we demonstrate that our method performs better than the Gaussian and Gaussian copula methods when these distributional assumptions are not satisfied. The method is applied to a breast cancer MicroRNA dataset to construct a gene network, which shows better interpretability than the Gaussian graphical model.

E0811: **Dimension reduction and data visualization for Frechet regression**
Presenter: **Qi Zhang**, The Pennsylvania State University, United States
Co-authors: Bing Li, Lingzhou Xue

With the rapid development of data collection techniques, complex data objects that are not in the Euclidean space are frequently encountered in new statistical applications. Frechet regression model provides a promising framework for regression analysis with metric space-valued responses, such as univariate probability distributions, covariance matrices and spherical data. We introduce a flexible sufficient dimension reduction method for Frechet regression to achieve two purposes: to mitigate the curse of dimensionality caused by the high-dimensional predictor, and to provide a tool for data visualization for Frechet regression. The approach is flexible enough to turn any existing SDR method for Euclidean (X,Y) into one for Euclidean X and metric space-valued Y. We derive the consistency and asymptotic convergence rate of the proposed methods. The finite-sample performance of the novel methods is illustrated through simulation studies of several commonly encountered metric spaces that include Wasserstein space, the space of symmetric positive definite matrices and the sphere. We illustrated the data visualization aspect of our method by the human mortality distribution data from the United Nation Databases and stroke data with CT hematoma densities.

E0388: **Benchpress: A scalable, platform-independent workflow to benchmark structure learning algorithms for graphical models**
Presenter: **Felix Rios**, University of Basel, Switzerland
Co-authors: Giusi Moffa, Jack Kuipers

Probabilistic graphical models (PGMs) play a central role in statistical data analysis, thanks to their compact and elegant way to represent complex dependence structures in multivariate probability distributions. In many realistic situations, ranging from disciplines such as social sciences to epidemiology, medicine, and biology, researchers are interested in finding the structure of an underlying model. Structure learning of PGMs is a computationally intensive task, with many and varied algorithms in constant development. The fast-moving pace and the heterogeneity of state-of-the-art algorithms pose a practical challenge for many researchers who wish to choose the most suitable algorithm for their specific problem or compare the performance of a novel method to the current state of the art. We will present a novel snakemake workflow, benchpress, which provides a unified platform for reproducible and scalable benchmarking and execution of structure learning algorithms for PGMs. Benchpress provides a platform where researchers and practitioners can easily compare existing algorithms available in the public domain in their original implementation, as well as enable other researchers to reproduce their results easily. We will demonstrate some of the functionalities in several analysis scenarios, typical for researchers and data scientists. The source code and documentation are publicly available from http://github.com/felixleopoldo/benchpress.

EO392 Room Virtual R21 RECENT ADVANCES IN BAYESIAN MODELING AND COMPUTATION Chair: Marco Ferreira

E0538: **PIPS: Screening the discrepancy function**
Presenter: **Rui Paulo**, Universidade de Lisboa, Portugal
Co-authors: Pierre Barbillon, Anabel Forte

Screening traditionally refers to the problem of detecting active inputs in a computer model. We develop a methodology that applies to screening, but the main focus is on detecting active inputs not in the computer model itself but rather on the discrepancy function that is introduced to account for model inadequacy when linking the computer model with field observations. We contend this is an important problem as it informs the modeller which are the inputs that are potentially being mishandled in the model, but also along which directions it may be less recommendable to use the model for prediction. The methodology is Bayesian and is inspired by the continuous spike and slab prior popularised by the literature on Bayesian variable selection. In our approach, and in contrast with previous proposals, a single MCMC sample from the full model allows us to compute the posterior probabilities of all the competing models, resulting in a methodology that is computationally very fast. The approach hinges on the ability to obtain posterior inclusion probabilities of the inputs, which are very intuitive and easy to interpret quantities, as the basis for selecting active inputs. For that reason, we name the methodology PIPS - posterior inclusion probability screening.

E0609: **Dynamic clustering of time series data with dynamically changing memberships**
Presenter: **Thais C O Fonseca**, Universidade Federal do Rio de Janeiro, Brazil
Co-authors: Victhor Sartorio

A new method is presented for clustering multivariate time-series based on Dynamic Linear Models. Whereas usual time-series clustering methods obtain static membership parameters, our proposal allows each time-series to dynamically change their cluster memberships over time. A mixture model is assumed for the time series and a flexible Dirichlet evolution for the mixture weights allows for smooth membership changes over time. Posterior estimates and predictions can be obtained through Gibbs sampling, but a more efficient method for obtaining point estimates is presented, based on Stochastic Expectation-Maximization and Gradient Descent. Finally, two applications illustrate the usefulness of our proposed model to

model both univariate and multivariate time-series: World Bank indicators for the renewable energy consumption of EU nations and the famous Gapminder dataset containing life-expectancy and GDP per capita for various countries.

E1027: **Predicting competitions by combining conditional logistic regression and subjective Bayes: An Academy Awards case study**
Presenter: **Christopher Franck**, Virginia Tech, United States
Co-authors: Christopher Wilson

Predicting the outcome of elections, sporting events, entertainment awards, and other competitions has long captured the human imagination. Such prediction is growing in sophistication in these areas, especially in the rapidly growing field of data-driven journalism intended for a general audience. Providing statistical methodology to probabilistically predict competition outcomes faces two main challenges. First, a suitably general modeling approach is necessary to assign probabilities to competitors. Second, the modeling framework must be able to accommodate expert opinion, which is usually available but difficult to fully encapsulate in typical data sets. We describe a recent effort to furnish statistical methodology that (i) overcomes both challenges, and (ii) is also of interest to the broad audience served by data journalists. The analysis of 2019 and 2020 Academy Awards data provides a case study, and we will also discuss the opportunities and challenges faced by statisticians and data journalists who embark on these sorts of collaborations.

E1046: **Objective Bayesian model selection for spatial hierarchical models with intrinsic conditional autoregressive priors**
Presenter: **Marco Ferreira**, Virginia Tech, United States
Co-authors: Erica Porter, Christopher Franck

Bayesian model selection is developed via fractional Bayes factors to simultaneously assess spatial dependence and select regressors in Gaussian hierarchical models with intrinsic conditional autoregressive (ICAR) spatial random effects. Selection of covariates and spatial model structure is difficult, as spatial confounding creates a tension between fixed and spatial random effects. The use of fractional Bayes factors allows for the selection of fixed effects and spatial model structure under automatic reference priors for model parameters, which obviates the need to specify hyperparameters for priors. We also derive the minimal training size for the fractional Bayes factor applied to the ICAR model under the reference prior. We perform a simulation study to assess the performance of our approach and we compare results to other readily available methods. We demonstrate that our fractional Bayes factor approach assigns a low posterior model probability to spatial models when data is truly independent and reliably selects the correct covariate structure with the highest probability within the model space. Finally, we demonstrate our Bayesian model selection approach with applications to county-level median household income in the contiguous United States and residential crime rates in the neighborhoods of Columbus, Ohio.

EO058 Room Virtual R22 RECENT DEVELOPMENTS IN SPATIAL STATISTICS Chair: Soutir Bandyopadhyay

E0213: **Modeling massive multivariate spatial data with the basis graphical lasso**
Presenter: **William Kleiber**, University of Colorado, United States
Co-authors: Mitchell Krock, Dorit Hammerling, Stephen Becker

A new modeling framework is proposed for highly multivariate spatial processes that synthesizes ideas from recent multiscale and spectral approaches with graphical models. The basis graphical lasso writes a univariate Gaussian process as a linear combination of basis functions weighted with entries of a Gaussian graphical vector whose graph is estimated from optimizing an L_1 penalized likelihood. We extend the setting to a multivariate Gaussian process where the basis functions are weighted with Gaussian graphical vectors. We motivate a model where the basis functions represent different levels of resolution, and the graphical vectors for each level are assumed to be independent. Using an orthogonal basis grants linear complexity and memory usage in the number of spatial locations, the number of basis functions, and the number of realizations. An additional fusion penalty encourages a parsimonious conditional independence structure in the multilevel graphical model. We illustrate our method on a large climate ensemble from the National Center for Atmospheric Research's Community Atmosphere Model that involves 40 spatial processes.

E0215: **Flexible and fast spatial return level estimation via a spatially-fused penalty**
Presenter: **Bo Li**, University of Illinois at Urbana-Champaign, United States
Co-authors: Danielle Sass, Brian Reich

Spatial extremes are common for climate data as the observations are usually referenced by geographic locations and dependent when they are nearby. An important goal of extremes modeling is to estimate the T -year return level. Among the methods suitable for modeling spatial extremes, perhaps the simplest and fastest approach is the spatial generalized extreme value (GEV) distribution and the spatial generalized Pareto distribution (GPD) that assume marginal independence and only account for dependence through the parameters. Despite the simplicity, simulations have shown that return level estimation using the spatial GEV, and spatial GPD still provides satisfactory results compared to max-stable processes, which are asymptotically justified models capable of representing spatial dependence among extremes. However, the linear functions used to model the spatially varying coefficients are restrictive and may be violated. We propose a flexible and fast approach based on the spatial GEV and spatial GPD by introducing fused lasso and fused ridge penalty for parameter regularization. This enables improved return level estimation for large spatial extremes compared to the existing methods.

E0216: **Computational developments and applications of the multi-resolution approximation for massive spatial data**
Presenter: **Lewis Blake**, Colorado School of Mines, United States
Co-authors: Huang Huang, Matthias Katzfuss, Dorit Hammerling

Recent developments of computational implementations and applications of the Multi-Resolution Approximation (MRA) spatial model are presented. Prediction and parameter inference are performed via a basis function representation of a Gaussian Process. Firstly, we introduce a parallel version of the MRA in Matlab for distributed computing environments. This implementation builds upon and improves a previous approach, facilitating computations with data sets on the order of 50 million observations while reducing execution times by up to 75%. Secondly, we provide a first glance at extending the MRA to model nonstationarity at a global scale. A basis function representation of covariance function parameters allows us to capture nuanced spatial dependence across large domains. The extension is natural for two reasons: (i) the MRA has been shown to be one of the most computationally efficient and accurate models to analyze large spatial data sets, (ii) the hierarchical domain partitioning imposed by the MRA explicitly defines regions of spatial dependence allowing for careful construction of nonstationary covariance functions. We apply this methodology to global measurements of sea surface temperature data from the Moderate Resolution Imaging Spectroradiometer onboard NASA's Aqua satellite.

E0217: **Adapting conditional simulation using circulant embedding for irregularly spaced data**
Presenter: **Soutir Bandyopadhyay**, Colorado School of Mines, United States
Co-authors: Maggie Bailey, Douglas Nychka

Computing an ensemble of random fields using conditional simulation is an ideal method for retrieving accurate estimates of a field conditioned on available data and for quantifying the uncertainty of these realizations. However, methods for generating random realizations are computationally demanding, especially when the estimates are conditioned on numerous observed data and for large domains. A new and approximate conditional simulation approach is applied that uses circulant embedding, a fast method for simulating Gaussian processes. However, standard CE is restricted to simulating stationary Gaussian processes (possibly anisotropic) on regularly spaced grids. We explore two new algorithms that adapt CE for irregularly spaced data points with applications to the U.S. Geological Survey's software ShakeMap, which provides near-real-time maps of shaking

intensity after a significant earthquake occurs. It is found that one method provides better accuracy and efficiency.

EO458 Room Virtual R23 RANDOM MATRIX THEORY AND ITS APPLICATIONS Chair: Monika Bhattacharjee

E0505: **Random matrices with independent entries: Beyond non-crossing partitions**
Presenter: **Priyanka Sen**, Indian Statistical Institute, Kolkata, India

The scaled standard Wigner matrix is known to have the semi-circular distribution as its LSD. The moments of the LSD are described with the help of non-crossing pair partitions. There have been several extensions of this result. We shall discuss the LSD of symmetric matrices with independent entries under certain moment assumptions and find a suitable set of partitions that help describe the moments of the LSD. The set of partitions found is in general larger than the class of non-crossing partitions and pose some interesting combinatorial questions. The LSD result will also help us bring several existing LSD results under one umbrella. For example, results on the standard Wigner matrix, the adjacency matrix of a sparse homogeneous Erdos-Renyi graph, heavy-tailed Wigner matrix, some banded Wigner matrices, Wigner matrices with variance profile follow from our result.

E0823: **Asymptotics 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. Under quite weak conditions, we prove, in a unified way, that the limiting spectral distribution (LSD) of any symmetric polynomial in the sample autocovariance matrices, after suitable centering and scaling, exists and is nondegenerate. We use methods from free probability in conjunction with the method of moments to establish our results. In addition, we are able to provide a general description of the limits in terms of some freely independent variables. We also establish asymptotic normality results for the traces of these matrices. We suggest statistical uses of these results in problems such as order determination of high dimensional MA and AR processes and testing of hypotheses for coefficient matrices of such processes.

E0920: **High dimensional multiplier bootstrap tests with applications to MANOVA**
Presenter: **Nilanjan Chakraborty**, Michigan State University, United States

The focus is on the Gaussian approximation of normalized sums of high dimensional random vectors over a class of convex sets. The class is flexible and general enough as it is defined through intersections of half-spaces and parametrized by a class of matrices. This newly proposed convex class helps us to quantify the effect of sparsity on the explicit convergence rate of the quality of the approximation. The resulting novel multiplier bootstrap method over this class allows conducting MANOVA tests for high-dimensional means. The resulting test is distribution- and correlation-free, and it can also be applied for Linear Hypothesis testing of means under a high-dimensional setup. The simulation studies for size and power conducted under different settings demonstrate the superiority of our approach over the available methods.

E0950: **Fluctuations of linear eigenvalue statistics of time dependent random circulant matrices**
Presenter: **Shambhu Nath Maurya**, Indian Institute of Technology Bombay, India

Fluctuations of linear eigenvalue statistics of random circulant matrices are discussed when the entries are independent Brownian motion. With polynomial test functions, we discuss the joint fluctuation and tightness of the time-dependent linear eigenvalue statistics of these matrices as the dimension of matrices goes to infinite. We see that the limit law is a Gaussian process with a nice variance structure. The methods of proofs are mainly combinatorial, based on some results of process convergence, trace formula of circulant matrix, method of moments and Wick's formula. This method can be applied to study fluctuations of linear eigenvalue statistics of other patterned random matrices, namely; Toeplitz, Hankel, reverse circulant and symmetric circulant matrices.

EO549 Room Virtual R24 FUNCTIONAL DATA ANALYSIS AND HIGH-DIMENSIONAL STATISTICS Chair: Michelle Carey

E1516: **Sequential detection of emergent phenomena within functional data**
Presenter: **Edward Austin**, Lancaster University, United Kingdom
Co-authors: Idris Eckley, Lawrence Bardwell

Detecting anomalies in a sequential setting is a well-studied area of research, however, the sequential detection of anomalies within partially observed functional data, termed emergent anomalies, is an open problem. Classical sequential detection approaches look for changes in the parameters, or structure, of point data and are not equipped to handle the complex nonstationarity and dependency structure of functional data. Existing functional data approaches, on the other hand, require the full observation of the curve before anomaly detection can take place. Motivated by an application arising from telecommunication engineering, we propose a new method that performs sequential detection of anomalies in partially observed functional data. The new method, called FAST, captures the common shape of the curves using Principal Differential Analysis and uses a form of CUSUM test to monitor a new functional observation as it emerges. The performance of FAST is then assessed on both simulated data and telecommunications data, demonstrating the effectiveness of the test in a range of settings.

E1552: **Functional data analysis of three-dimensional surface data: Examples on human face**
Presenter: **Stanislav Katina**, Masaryk Univesity, Czech Republic

The advent of high-resolution imaging has made data on surface shape widespread. Methods for the analysis of shape based on landmarks are well established, but high-resolution data require a functional approach. The starting point is a systematic and consistent description of each surface shape (using, e.g., curves, and surface patches) and a method for creating this is described. Two innovative forms of analysis are introduced. The first uses surface integration to address issues of registration, principal component (PC) analysis, all in functional form. Computational issues are handled through discrete approximations to integrals based on appropriate surface area weighted sums. The second innovation is to focus on subspaces where interesting behaviour such as group differences are exhibited, rather than on individual PCs. All these ideas are developed and illustrated in the important context of the human facial shape of patients before and after orthognathic surgery and patients with psychotic or other disorders and controls, with a strong emphasis on effective visual communication of effects of interest. All presented methods are implemented in R as part of a face3d package development.

E1429: **Finite-sample exact prediction bands for functional data based on conformal prediction**
Presenter: **Matteo Fontana**, European Commission, Joint Research Centre, Italy
Co-authors: Simone Vantini, Jacopo Diquigiovanni

The focus is on the prediction of a new unobserved functional datum given a set of observed functional data, possibly in presence of covariates, either scalar, categorical, or functional. In particular, we will present an approach (i) able to provide prediction regions that could be visualized in the form of bands, (ii) guaranteed with exact coverage probability for any sample size, (iii) not relying on parametric assumptions about the specific distribution of the functional data set, and finally (iv) being computationally efficient. The method is built on a combination of ideas coming from the recent literature pertaining to functional data analysis (i.e., the statistical analysis of datasets made of functions) and conformal prediction (i.e., a nonparametric predictive approach from Machine Learning). We will present the general theoretical framework and some simulations enlightening the flexibility of the approach and the effect on the amplitude of prediction bands of different algorithmic choices.

E1747: **Interpretable discriminant analysis for functional data supported on general random domains**
Presenter: **Eardi Lila**, University of Washington, United States

A novel framework is introduced for the classification of functional data supported on non-linear, and possibly random, manifold domains. 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 into a regularized multivariate functional linear regression model. This allows us to directly estimate the discriminant direction while controlling for its complexity with appropriate differential regularizations. The approach does not require prior estimation of the covariance structure of the functional predictors, which is computationally not feasible in our application setting. We apply the proposed method to the Alzheimer's Disease Neuroimaging Initiative data and are able to estimate novel discriminant directions that capture both geometric and thickness predictive features of the cerebral cortex.

EO661 Room Virtual R25 RECENT DEVELOPMENTS IN ROBUST METHODOLOGY Chair: Peter Rousseeuw

E0330: **A cellwise robust lasso estimator**
Presenter: **Ines Wilms**, Maastricht University, Netherlands
Co-authors: Lea Bottmer, Christophe Croux

The high-dimensional multiple regression model is an important workhorse for data scientists. The lasso is a popular estimator to reduce the dimensionality by imposing sparsity on the estimated regression parameters. The lasso is, however, not a robust estimator. Nevertheless, outliers frequently occur in high-dimensional datasets. We propose the sparse shooting *S*, a cellwise robust lasso estimator. The resulting regression coefficients are sparse, meaning that many of them are set to zero, hereby selecting the most relevant predictors. As such, the sparse shooting *S* is computable in high-dimensional settings with more predictors than observations. Moreover, a distinct feature of this estimator is its ability to deal with cellwise contamination, where many cells of the design matrix of the predictor variables may be outlying. We compare its performance to several other sparse and/or robust regression estimators.

E0807: **Robust clustering with cellwise trimming**
Presenter: **Luis Angel Garcia-Escudero**, Universidad de Valladolid, Spain
Co-authors: Diego Rivera-Garcia , Agustin Mayo-Iscar, Joaquin Ortega

It is known that a small fraction of outlying measurements can harmfully affect classical Cluster Analysis techniques. Trimming is a simple and sensible procedure to achieve robustness in statistical procedures. Some procedures have been introduced in clustering that allows trimming complete observations. However, trimming entire observations, rather than just trimming the most outlying cells, can be too extreme at sacrificing a lot of valuable information. This is especially the case when dimension increases because many observations being completely free of outlying cells are sometimes difficult to expect. In order to deal with this problem, a cellwise trimming approach based on affine subspace approximations and robust regression techniques is introduced. The approach is particularized to functional clustering problems.

E1256: **Robust correspondence analysis**
Presenter: **Marco Riani**, University of Parma, Italy
Co-authors: Anthony Atkinson, Aldo Corbellini, Francesca Torti

Correspondence analysis is a method for displaying information from two-way tables of count data. Typically, the rows are subjects (in our first example the 28 countries of the European Union) and the columns are response categories (in that case the cost range of clothes). The main result is a two-dimensional plot showing the structure of the data. The theory and practice of correspondence analysis are presented in several books by Greenacre. Little attention seems to have been given to the effect of outliers on correspondence analysis nor to the desirability and practice of robust estimation. We introduce a robust form of correspondence analysis based on minimum covariance determinant estimation. This leads to the systematic deletion of outlying rows of the table and to plots of greatly increased informativeness. Our examples are trade flows of clothes and consumer evaluations of the perceived properties of cars. The robust method requires that a specified proportion of the data be used in fitting. To accommodate this requirement we provide an algorithm that uses a subset of complete rows and one row partially, both sets of rows being chosen robustly. We prove the convergence of this algorithm.

E1343: **Transforming variables to central normality**
Presenter: **Jakob Raymaekers**, KULeuven, Belgium
Co-authors: Peter Rousseeuw

Many real data sets contain numerical features (variables) whose distribution is far from normal (Gaussian). Instead, their distribution is often skewed. In order to handle such data, it is customary to preprocess the variables to make them more normal. The Box-Cox and Yeo-Johnson transformations are well-known tools for this. However, the standard maximum likelihood estimator of their transformation parameter is highly sensitive to outliers, and will often try to move outliers inward at the expense of the normality of the central part of the data. We propose a modification of these transformations as well as an estimator of the transformation parameter that is robust to outliers, so the transformed data can be approximately normal in the center and a few outliers may deviate from it. It compares favorably to existing techniques in an extensive simulation study and on real data. An implementation of the proposed method is available in the R package cellWise.

EO700 Room Virtual R26 BAYESIAN NONPARAMETRIC MODELS Chair: Luis Gutierrez

E0603: **Generalized additive neutral to the right regression for survival analysis**
Presenter: **Alan Riva-Palacio**, Universidad Nacional Autonoma de Maxico, Mexico

A novel Bayesian nonparametric model is presented for regression in survival analysis. The model builds on the neutral to the right model and on the Cox proportional hazards model. The use of a vector of dependent Bayesian nonparametric priors allows us to efficiently model the hazard as a function of covariates whilst allowing non-proportionality. Properties of the model and inference schemes will be discussed. The method will be illustrated using simulated and real data.

E0689: **A new flexible Bayesian hypothesis test for multivariate data**
Presenter: **Ivan Gutierrez**, Pontificia Universidad Catolica de Chile, Chile
Co-authors: Danilo Alvares, Luis Gutierrez

A Bayesian hypothesis testing procedure is proposed for comparing the multivariate distributions of several treatment groups against a control group. This test is based on a flexible model for the group distributions with the introduction of a random binary vector such that if its jth element equals one, then the jth treatment group is merged with the control group. The group distributions' flexibility comes from a dependent Dirichlet process, while the prior distribution from the latent vector ensures a multiplicity correction to the testing procedure. We explore the posterior consistency of the Bayes factor and provide a Monte Carlo simulation study comparing the performance of this procedure with state-of-the-art alternatives. The results show that the presented method performs better than competing approaches. Finally, we apply our proposal to two classical experiments. The first one studies the effects of tuberculosis vaccines on multiple health outcomes for rabbits, and the second one analyzes the effects of two drugs on weight gains for rats. In both applications, we find relevant differences between the control group and at least one treatment group.

E0885: **Bayesian analysis of jointly heavy-tailed data**
Presenter: **Karla Vianey Palacios Ramirez**, University of Edinburgh, United Kingdom

A novel flexible Bayesian nonparametric model is introduced for a jointly heavy-tailed response. The proposed model induces a prior on the space

of multivariate distributions with heavy-tailed marginal distributions. The proposal resorts to an infinite mixture of multivariate Erlang distributions, with a specific parameterization, to learn about the body and joint tails of a jointly heavy-tailed distribution. A predictor dependent version of the model is devised to learn about the effect of covariates on a jointly heavy-tailed response. Sufficient conditions to induce a multivariate heavy-tailed distribution are studied, and posterior inference following a Bayesian nonparametric approach is developed. The simulation study suggests that the proposed method performs well in a variety of scenarios. We showcase the application of the proposed methods in a case study in neuroscience.

E0880: **Clustering point patterns on linear networks**
Presenter: **Ramses Mena**, UNAM, Mexico

An unsupervised classification method for point events occurring on a network of lines is proposed. The idea relies on the distributional flexibility and practicality of Bayesian nonparametric mixture models to discover the clustering structure featuring observations from a particular phenomenon taking place on a given set of edges. By incorporating the spatial effect in the random partition distribution, induced by a Dirichlet process, one is able to control the distance between edges and events, thus leading to an appealing method to point clustering. A Gibbs sampler algorithm is proposed and evaluated with a sensitivity analysis. The proposal is motivated and illustrated by the analysis of crime and violence patterns in Mexico City.

EO106 Room Virtual R27 MODERN STATISTICAL METHODS IN DATA SCIENCE Chair: Yichuan Zhao

E0299: **Optimal classification for functional data**
Presenter: **Guanqun Cao**, Auburn University, United States

A central topic in functional data analysis is how to design an optimal decision rule, based on training samples, to classify a data function. We exploit the optimal classification problem when data functions are Gaussian processes. Sharp nonasymptotic convergence rates for minimax excess misclassification risk are derived in both settings that data functions are fully observed and discretely observed. We explore two easily implementable classifiers based on discriminant analysis and deep neural network, respectively, which are both proven to achieve optimality in a Gaussian setting. Our deep neural network classifier is new in literature, demonstrating outstanding performance even when data functions are non-Gaussian. In the case of discretely observed data, we discover a novel critical sampling frequency that governs the sharp convergence rates. The proposed classifiers perform favorably in finite-sample applications, as we demonstrate through comparisons with other functional classifiers in simulations and one real data application.

E0714: **Robust prediction of survival outcomes through unified Bayesian analysis of nonparametric transformation models**
Presenter: **Catherine Liu**, The Hong Kong Polytechnic University, Hong Kong

The linear transformation model is a general semiparametric survival model that includes the widely-used Cox model, proportional odds model, accelerated failure time model, and beyond. It is known to be challenging to predict the survival outcomes when the transformation function and the distribution of the error term are both unspecified owning to model identifiability. Unlike the strategy of identifying the model first and then fitting the model in existing Bayesian literature, we develop a unified Bayesian procedure to estimate the nonparametric functions and the parametric component simultaneously and jointly under quite mild assumptions after exponential transformation. We construct an augmented Gamma process smoothed by I-spline functions as the prior for the monotonic transformation function. We refine the parametric estimator by a posterior projection owing to the constraints of identifiability. The predicted conditional survival function behaves very well whatever the underlying true models are in numerical analysis. The MCMC sampler is built based on the No-U-Turn sampler (NUTS) by Stan. Comprehensive simulations and an application to real data illustrate the methodology broadly using an R packageBuLTM.

E0870: **Learning privately over distributed features: An ADMM sharing approach**
Presenter: **Peng Liu**, University of Kent, United Kingdom

Distributed machine learning has been widely studied in order to handle exploding amount of data. We study an important yet less visited distributed learning problem where features are inherently distributed or vertically partitioned among multiple parties, and sharing of raw data or model parameters among parties is prohibited due to privacy concerns. We propose an ADMM sharing framework to approach risk minimization over distributed features, where each party only needs to share a single value for each sample in the training process, thus minimizing the data leakage risk. We establish convergence and iteration complexity results for the proposed parallel ADMM algorithm under non-convex loss. We further introduce a novel differentially private ADMM sharing algorithm and bound the privacy guarantee with carefully designed noise perturbation. The experiments based on a prototype system show that the proposed ADMM algorithms converge efficiently in a robust fashion, demonstrating an advantage over gradient-based methods especially for data set with high dimensional feature spaces.

E1583: **How to deal with missing values in the high dimensional supervised context**
Presenter: **Hadrien Lorenzo**, INRIA, France
Co-authors: Olivier Cloarec, Jerome Saracco

The exploratory analysis searches for a data structure hidden from the user. In the supervised case, this data structure may only represent a very small portion of the total data structure, which is all the more volatile as the sampling is small or the quantity of descriptors is large. The management of missing data is classically done without taking into account the supervised nature of the research question. This assumes that the structure associated with the answer is accessible to an unsupervised method. This assumption is questionable, especially in high dimensions. Indeed, in high dimensions, an unsupervised approach of imputation estimates on variables is useless to the prediction model. In practice, only the initialization values of the imputation (often iterative algorithms of EM type) are kept for the variables of interest. A solution to the problem of supervised imputation in high dimensions is presented. The context is the linear model solved with the PLS (Partial Least Squares) method. The construction of subspaces in the linear context makes it possible to "skip" the missing data, what does the Nipals method. On the other hand, the regularized PCA approach, through multiple imputation implemented in the missMDA R-package, allows the estimation of missing data for unsupervised problems. The objective is to compare these different approaches in contexts more or less favorable to each of them.

EO216 Room Virtual R28 STOCHASTIC MODELS FOR DEPENDENCE Chair: Sebastian Fuchs

E1172: **Markov product invariance in classes of bivariate copulas characterized by univariate functions**
Presenter: **Marco Tschimpke**, Paris Lodron University Salzburg, Austria
Co-authors: Wolfgang Trutschnig, Juan Fernandez Sanchez

The aim is to extend and sharpen some results concerning the notion of Markov product idempotence in some well-known classes of copulas. Focusing on families of copulas which are characterized by univariate functions we show that in the class of extreme-value copulas, in the class of diagonal copulas and in some special class of copulas represented by measure-preserving transformations only the usual suspects (if contained in the class) are idempotent, namely the product copula Π and minimum copula M. Additionally, we prove the conjecture that the only idempotent Archimedean copula is the product copula Π.

E1097: **Asymmetric dependence modeling for contingency table with an ordinal dependent variable**
Presenter: **Zheng Wei**, University of Maine, United States
Co-authors: Daeyoung Kim, Li Wang

For the analysis of a contingency table with an ordinal dependent variable, a subcopula based asymmetric association measure is developed. The

subcopula regression-based association measure exploits the subcopula regression to quantify the strength of the association structure in a model-free manner. Unlike the existing measures of asymmetric association, the subcopula-based measure is insensitive to the number of categories in a variable, and thus, the magnitude of the proposed measure can be interpreted as the degree of asymmetric association in the contingency table. The methodology consists of subcopula score, subcopula regression, subcopula regression-based association measure and its decompositions. The sequential decompositions of the proposed association measure evaluate the contribution of the subsets of independent variables to the overall association in various forms.

E1187: **Kendall's tau estimator for zero-inflated discrete distributions**
Presenter: **Elisa Perrone**, Eindhoven University of Technology, Netherlands
Co-authors: Zhuozhao Zhan, Edwin van den Heuvel

Zero-inflated data naturally appears in many applications such as health care, weather forecasting, and insurance. Analyzing zero-inflated data is challenging as the high amount of observations in zero invalidates standard statistical techniques. For example, assessing the level of dependence between two zero-inflated random variables becomes difficult due to limitations when applying standard rank-based measures of association, such as Kendall's tau or Spearman's rho. Recent work tackles this issue and suggests an estimator of Kendall's tau for zero-inflated continuous distributions. However, such an estimator does not show satisfactory performances for zero-inflated count data. We fill this gap and propose an adjusted estimator specific for zero-inflated discrete distributions. We derive the estimator analytically and show that it outperforms existing estimators in various simulated scenarios. Finally, we investigate the interpretability of the proposed estimator by studying its achievable bounds.

E1233: **High dimensional simulation of exchangeable Bernoulli distributions**
Presenter: **Roberto Fontana**, Politecnico di Torino, Italy
Co-authors: Patrizia Semeraro

High-dimensional simulation of exchangeable multivariate Bernoulli distributions is a challenging and important issue in applications, as for example in credit risk models. The main contribution is, even for high dimensions, algorithms to sample from exchangeable multivariate Bernoulli distributions and to determine the distributions and the bounds of a wide class of indices and measures of probability mass functions. Unlike the algorithms present in the literature the proposed method gives the possibility to simulate also from negatively correlated distributions. Such a method is based on the geometrical structure of the class of exchangeable Bernoulli probability mass functions, which are points in a convex polytope whose extremal points are analytically known. Estimation and testing are also briefly addressed.

EO631 Room Virtual R29 APPLIED BAYESIAN MODELS Chair: Emanuele Aliverti

E1079: **Variational approximation for stochastic volatility via continuous random walk processes**
Presenter: **Nicolas Bianco**, University of Padua, Italy
Co-authors: Mauro Bernardi

Discrete-time stochastic volatility (SV) represent a valid alternative to GARCH models, that are easy to estimate, but have relevant drawbacks. However, the formulation of the volatility as a latent process makes parameters estimation more complicated for the SV models. The estimation has been previously tackled within an approximate Bayesian framework by leveraging the variational Bayes approach. One approach locally approximates the underlying stochastic process, while another suggests using a multivariate Gaussian as an approximation of the joint distribution of the latent volatilities with a fixed structure of the variance-covariance matrix. We propose variational methods for parameters estimation and signal extraction in a Bayesian context that relies on a flexible approximation of the volatility through a continuous random walk (CRW) process. The latter is a semi-parametric process that is consistent with respect to the choice of the locations and it has a sparse precision matrix that enables efficient computations. Although the point estimates of the posterior volatility are precise, the CRW process is homoscedastic which is in contrast with the true behaviour of the volatility a posteriori. Therefore, to get accurate HPD intervals, we extend the CRW process to account for the heteroscedastic behaviour by allowing its scale to follow a process leading to a hierarchical random field approximation.

E1415: **A closed-form filter for binary time series**
Presenter: **Augusto Fasano**, Collegio Carlo Alberto and European Commission - Joint Research Centre, Italy
Co-authors: Giovanni Rebaudo, Daniele Durante, Sonia Petrone

Non-Gaussian state-space models arise in several applications, and within this framework, the binary time series setting provides a relevant example. However, unlike for Gaussian state-space models - where filtering, predictive and smoothing distributions are available in closed form - binary state-space models require approximations or sequential Monte Carlo strategies for inference and prediction. This is due to the apparent absence of conjugacy between the Gaussian states and the likelihood induced by the observation equation. We prove that the filtering, predictive and smoothing distributions in dynamic probit models with Gaussian state variables are, in fact, available and belong to a class of unified skew-normals (SUN) whose parameters can be updated recursively. Also, the key functionals of these distributions are, in principle, available, but their calculation requires the evaluation of multivariate Gaussian cumulative distribution functions. Leveraging SUN properties, we address this issue via novel Monte Carlo methods based on independent samples from the smoothing distribution, that can easily be adapted to the filtering and predictive case, thus improving state-of-the-art approximate and sequential Monte Carlo inference in small-to-moderate dimensional studies. Novel sequential Monte Carlo procedures are also developed to deal with online inference in high dimensions. Performance gains over competitors are outlined in a financial application.

E1225: **A generalized partial credit model for network dependent latent traits with an application on modeling students' ability**
Presenter: **Massimiliano Russo**, Harvard Medical School, United States

Partial credit models are widely used in item response theory to obtain interpretable inference when analyzing polytomous data. They map responses to items into continuous constructs that summarize individual latent traits. A main assumption of these models is that the latent traits are independent across subjects. This is a reasonable assumption when the measured individuals have few or no interactions. However, in some cases individual relationships are likely to be important predictors of the latent traits. For example, when monitoring students performances it is likely that high-achieving/low-achieving students bond together. Consequently, friends are likely to share similar latent trait values. To characterize the inter-individual dependence of the latent traits, we propose a novel generalized partial credit model that accounts for network connectivity patterns. Specifically, we rely on a conditional auto regressive (CAR) model for the distribution of an individual latent trait conditional to the others. The strength of the network dependence on the latent traits is directly quantified by the parameters of the CAR model, and can be tested from the data with a spike-and-slab prior. We illustrate the performance of the proposed model in simulations and in a real data application evaluating students' ability conditionally on their Facebook friendship network.

E1221: **A Bayesian approach for inference on probabilistic surveys**
Presenter: **Roberto Casarin**, University Ca' Foscari of Venice, Italy
Co-authors: Federico Bassetti, Marco Del Negro

A non-parametric Bayesian approach is proposed to the estimation of forecast densities in probabilistic surveys. We use it to study the evolution of the subjective forecast distribution for inflation from the U.S. Survey of Professional Forecasters over the past forty years.

EO800 Room Virtual R30 MATHEMATICAL AND STATISTICAL FOUNDATIONS FOR DEEP LEARNING Chair: Lizhen Lin

E1194: **Regression and doubly robust off-policy learning on low-dimensional manifolds by neural networks**
Presenter: **Wenjing Liao**, Georgia Institute of Technology, United States

Many data in real-world applications are in a high-dimensional space but exhibit low-dimensional structures. In mathematics, these data can be modeled as random samples on a low-dimensional manifold. The goal is to estimate a target function or learn an optimal policy using neural networks. The basis is an efficient approximation theory of deep ReLU networks for functions supported on a low-dimensional manifold. We further establish the sample complexity for regression and off-policy learning with finite samples of data. When data are sampled on a low-dimensional manifold, the sample complexity crucially depends on the intrinsic dimension of the manifold instead of the ambient dimension of the data. These results demonstrate that deep neural networks are adaptive to low-dimensional geometric structures of data sets.

E1228: **Deep generative models for nonparametric estimation of a singular distribution**
Presenter: **Minwoo Chae**, Pohang University of Science and Technology, Korea, South
Co-authors: Lizhen Lin, Yongdai Kim, Dongha Kim

While deep generative models are popularly used to model high-dimensional data, there is a lack of mathematical understanding. We investigate the statistical properties of deep generative models from a nonparametric distribution estimation viewpoint. In the considered model, data are assumed to concentrate around some low-dimensional structure. Estimating the distribution supported on this low-dimensional structure, such as a low-dimensional manifold, is challenging due to its singularity with respect to the Lebesgue measure in the ambient space. In particular, a likelihood approach can fail to estimate the target distribution consistently. We obtain convergence rates with respect to the Wasserstein metric for two methods: a sieve MLE based on the perturbed data and a GAN type estimator. Our analysis gives some insights into i) how deep generative models can avoid the curse of dimensionality, ii) how likelihood approaches work for singular distribution estimation, and iii) why GAN performs better than likelihood approaches.

E1373: **Adaptive variational deep learning**
Presenter: **Ilsang Ohn**, University of Notre Dame, Korea, South
Co-authors: Lizhen Lin

A novel variational Bayes method is introduced for deep neural networks, which we call adaptive variational deep learning. The proposed method first obtains individual variational posterior over deep neural network models with varying network width (i.e., number of hidden nodes per layer) and combines them with certain weights to obtain a variational posterior over the entire deep neural network model. We show that the resulting variational posterior can obtain adaptive optimal contraction rates in a large number of statistical problems. Simulation results demonstrate that the strong empirical performance of the adaptive variational deep learning.

E1755: **Adaptivity of deep learning to intrinsic dimensionality via mixed and anisotropic smoothness for high dimensional input**
Presenter: **Taiji Suzuki**, University of Tokyo / RIKEN-AIP, Japan
Co-authors: Atsushi Nitanda, Sho Okumoto

The estimation error of deep learning for high dimensional input, which is possibly infinite-dimensional, is discussed. Deep learning has been applied to high dimensional data such as image and voice signals. However, a usual nonparametric regression analysis asserts that a nonparametric method such as deep learning could suffer from the curse of dimensionality. We show that deep learning can extract informative features from the high dimensional input and avoid the curse of dimensionality if the true function has so-called mixed and anisotropic smoothness. Moreover, we also investigate convolutional neural networks and show that the curse of dimensionality can be avoided even though the input is infinite-dimensional. In particular, we show that the dilated convolution is advantageous when the smoothness of the target function has a sparse structure.

EO136 Room Virtual R34 MARKOV SWITCHING MODELS Chair: Maddalena Cavicchioli

E1248: **Markov switching positive conditional mean models: Structure and examples**
Presenter: **Abdelhakim Aknouche**, Qassim University, Saudi Arabia
Co-authors: Christian Francq

A Markov-Switching autoregressive conditional mean model, valued in the set of nonnegative numbers, is considered. The conditional distribution of this model is a finite mixture of nonnegative distributions whose conditional mean follows a GARCH-like pattern with parameters depending on the state of a Markov chain. Three different variants of the model are examined depending on how the lagged values of the mixing variable are integrated into the conditional mean equation. The model includes, in particular, Markov mixture versions of various well-known nonnegative time series models such as the autoregressive conditional duration (ACD) model, the integer-valued GARCH (INGARCH) model, and the Beta observation-driven model. Under contraction in mean conditions, it is shown that the three variants of the model are stationary and ergodic when the stochastic order and the mean order of the mixing distributions are equal. The proposed conditions match those already known for Markov-switching GARCH models. We also give conditions for finite marginal moments. Applications to various mixture and Markov mixture count, duration, and proportion models are provided.

E0311: **Jointly determining the state dimension and lag order for Markov-switching vector autoregressive models**
Presenter: **Nan Li**, Barclays, United Kingdom
Co-authors: Simon Kwok

The focus is on the problem of joint selection of the state dimension and lag order for a class of Markov-switching vector autoregressive models, in which all parameters are presumed to be regime-dependent. To this end, three complexity-penalized criteria are considered, and a new criterion is derived by minimising the Kullback-Leibler divergence. The accuracy of the procedure is evaluated by means of Monte Carlo experiments. We illustrate the usefulness of the joint model selection procedure with empirical applications to the modelling of business cycles in the U.S. and Australia.

E0939: **On spectral analysis and forecasting performance of regime switching GARCH-MIDAS models**
Presenter: **Jie Cheng**, Keele University, United Kingdom

The general expressions of autoregressive functions and spectra for regime-switching GARCH-MIDAS models under certain conditions are constructed. Simulation studies of theoretical spectral density functions of these models are presented. The results show the different effect of switching characteristic, long-term and short-term GARCH components and how it governs the distributions of the volatility. These results can be regarded as a starting point to propose some specification tests. We also extend a previous result on forecast evaluation to the case of a regime-switching GARCH-MIDAS model and the theoretical results provide a new insight to forecasting performance evaluation, especially for regime-switching models.

E1074: **The Beveridge-Nelson decomposition of Markov switching (co)integrated time series**
Presenter: **Maddalena Cavicchioli**, University of Modena and Reggio Emilia, Italy

The aim is to derive the Beveridge-Nelson (BN) decomposition and the state space representation for various multivariate (co)integrated time series subject to Markov switching in regime. Then, we provide explicit expressions for the BN trend and cyclical components in terms of the matrices involved in the state space representation of the considered process. This gives computational advantages as only quantities already

available are involved in the calculations. Moreover, our matrix expressions in closed-form improve computational performance since they are readily programmable and greatly reduce the computational cost. Then we develop impulse-response function analysis and represent the BN trend component as a random walk. A numerical example illustrates the feasibility of the proposed approach.

EO521 Room Virtual R36 NETWORK STATISTICS Chair: Tracy Ke

E1433: **Model selection on random network models**
Presenter: **Swati Chandna**, Birkbeck, University of London, United Kingdom

Network summaries, such as the degree distribution, transitivity or average path length, have long been of interest to practitioners in applications ranging from social sciences to neuroscience. To parameterize a model in terms of a given set of network summaries of interest is a challenging task. Likewise, whether a chosen model is suitable for network summaries of interest is also not always clear. More generally, there is a need for statistical network models which can not only be fitted with the available computational constraints but are also appropriate for a specified set of network summaries of interest. For example, the stochastic block model does not parameterize average path length (the typical number of edges between pairs of nodes) but is a flexible model which can be easily fitted via standard least squares or likelihood techniques. We propose a general Bayesian framework for model selection on, in general, non-nested random network models. The criterion is formulated via Bayes factors and penalizing using the most widely used loss functions. The objective is to identify random network models such that a reasonable trade-off between the network features of interest and the complexity of the model is preserved.

E1653: **Tractably modelling dependence in networks beyond exchangeability**
Presenter: **Weichi Wu**, Tsinghua University, China
Co-authors: Sofia Olhede, Patrick Wolfe

A general framework is proposed for modelling network data that is designed to describe aspects of non-exchangeable networks. Conditional on latent (unobserved) variables, the edges of the network are generated by their finite growth history (with latent orders) while the marginal probabilities of the adjacency matrix are modeled by a generalization of a graph limit function (or a graphon). In particular, we study the estimation, clustering and degree behavior of the network in our setting. We determine (i) the minimax estimator of a composite graphon with respect to squared error loss; (ii) that spectral clustering is able to consistently detect the latent membership when the block-wise constant composite graphon is considered under additional conditions; and (iii) we are able to construct models with heavy-tailed empirical degrees under specific scenarios and parameter choices. This explores why and under which general conditions non-exchangeable network data can be described by a stochastic block model. The new modelling framework is able to capture empirically important characteristics of network data such as sparsity combined with heavy-tailed degree distribution, and add understanding as to what generative mechanisms will make them arise.

E1737: **The citation behavior of statisticians**
Presenter: **Jiashun Jin**, Carnegie Mellon University, United States

A data set has been collected and cleaned consisting of the bibtex and citation data of 83K papers published in 36 journals in statistics and related fields, spanning 41 years. The data set provides a great opportunity to study the citation behavior, trends, and patterns of statisticians. We are interested in (a) how to identify representative research topics in statistics, rank them, and use them to visualize the dissemination of ideas across different topics, (b) how to rank all the 36 journals, (c) how to identify the friendliest journal for a given topic, and (d) how to predict future citations and identify representative citation patterns. We propose to jointly model the bibtex and citation data by the Hofmann-Stigler model, and propose to use the TR-SCORE (among others) as a new approach to address these problems. A good understanding of this problem may help administrators in decision making, and individual authors for making plans for future research.

E1783: **The hierarchy of block models**
Presenter: **Majid Noroozi**, University of Memphis, United States
Co-authors: Marianna Pensky

There exist various types of network block models such as the Stochastic Block Model (SBM), the Degree Corrected Block Model (DCBM), and the Popularity Adjusted Block Model (PABM). While this leads to a variety of choices, the block models do not have a nested structure. In addition, there is a substantial jump in the number of parameters from the DCBM to the PABM. The objective is the formulation of a hierarchy of block model which does not rely on arbitrary identifiability conditions. We propose a Nested Block Model (NBM) that treats the SBM, the DCBM and the PABM as its particular cases with specific parameter values, and, in addition, allows a multitude of versions that are more complicated than DCBM but have fewer unknown parameters than the PABM. The latter allows one to carry out clustering and estimation without preliminary testing, to see which block model is really true.

EO118 Room Virtual R37 STATISTICAL METHODS FOR HIV RESEARCH Chair: Erica Moodie

E0585: **Viral dynamics as an outcome in HIV therapeutic vaccine trials: From AUC to dynamical modelling**
Presenter: **Rodolphe Thiebaut**, Bordeaux University U 1219 INSERM, France
Co-authors: Marie Alexandre, Melanie Prague

There is no HIV therapeutic vaccine available yet. The goal of such a vaccine is to allow patients immune systems for controlling viral replication without any antiretroviral therapy. Therefore, the evaluation of the candidates is performed through supervised antiretroviral treatment interruption (STI) where HIV infected patients are vaccinated while treated by antiretroviral treatment and then the HIV viral dynamics is studied during STI. However, antiretroviral treatments can be resumed quickly in case of viral replication for safety issues. The question is how to define and analyze the virological endpoint in such context. Viral load data are incomplete because of censoring due to a limit of detection or the censoring of the follow-up due to treatment resumption (drop-out). We recently proposed to compare the area under the outcome time curve (AUC) through a statistical test based on splines-based mixed-model accounting for both censoring and missingness mechanisms in the AUC estimation. The inferential properties were studied through a simulation study of a two-armed trial and compared to traditional methods. However, the limitation of the method in the context of a high level of drop-out underlines the relevance of replacing a descriptive model with a dynamical model based on ordinary differential equations defined by biological mechanisms that are potentially performing better predictions in such context.

E0978: **Counting viruses: Estimating the size of the latent HIV reservoir**
Presenter: **Michael Hudgens**, University of North Carolina, United States

Statistical methods will be discussed for quantifying the latent HIV reservoir in anti-retroviral therapy (ART) suppressed individuals. In particular, analysis will be considered of data from the quantitative virus outgrowth assay (QVOA), a type of serial limiting dilution assay which is used to estimate the number of infectious units per million (IUPM) resting CD4+ T cells. A simple, free publicly available R software package and web tool to analyze serial limiting dilution assay data will be described. Methods will also be considered for quantifying the size of the reservoir when additional viral RNA sequence data is available.

E0983: **Prediction of the time to treatment success from longitudinal left-censored viral load measurements**
Presenter: **Tarylee Reddy**, South African Medical Research Council, South Africa
Co-authors: Geert Molenberghs, Marc Aerts

Viral load measurements play a key role in monitoring antiretroviral treatment success in HIV positive patients. In this setting, treatment success is

formally defined as two consecutive viral load measurements less than 1000 copies/ml. Statistical challenges that arise in the analysis of longitudinal viral load measurements include: the handling of left-censored measurements and the non-linear evolution of viral load over time. We present a novel approach to estimate the time to treatment success, taking into account left-censoring and the biphasic evolution of viral load. In the first stage of the approach a mixed model, with random intercept and two random slopes is fitted to the data, where the partial information provided by the censored values is incorporated into the likelihood function. In the second stage, using the estimates from the mixed model, the probability of treatment success and the estimated time to treatment success is computed using a previous methodology. We apply the proposed methodology to an HIV/AIDS clinical trial, presenting the estimated time to treatment success and predicted viral load trajectory for selected patients.

E0164: **Value of information for HIV evidence synthesis**
Presenter: **Daniela De Angelis**, University of Cambridge, United Kingdom

Annual national estimates of the number of people living with HIV in England, particularly those who are unaware of their infection, have, for several years, been based on a Bayesian model that combines evidence from multiple sources of surveillance, survey data and prior beliefs. In such an evidence synthesis model it is important to know which parameters most affect the estimates and, therefore, the decision from the model; which of the parameter uncertainties drive the decision uncertainty; and what further data should be collected to reduce such uncertainty. These questions can be addressed by Value of Information (VoI) analysis, allowing estimation of the expected gain from learning specific parameters or collecting data of a given design. We introduce the concept of VoI for Bayesian evidence synthesis, using and extending ideas from health economics, computer modelling and Bayesian design. We then apply it to our HIV prevalence model. Results show which parameters contribute most to the uncertainty about each prevalence estimate, and the expected improvements in precision from specific amounts of additional data. These benefits can be traded with the costs of sampling to determine an optimal sample size.

EO579 Room Virtual R39 RECENT ADVANCES IN EXTREME RISK MEASURES ESTIMATION Chair: Antoine Usseglio-Carleve

E0199: **Semi-parametric estimation of multivariate extreme expectiles**
Presenter: **Elena Di Bernardino**, LJAD Laboratoire J.A. Dieudonna, France

The focus is on semi-parametric estimation of multivariate expectiles for extreme levels of risk. Multivariate expectiles and their extremes have been the focus of plentiful research in recent years. In particular, it has been noted that an alternative formulation of the underlying optimisation problem would be necessary due to the difficulty in estimating these values for elevated levels of risk, an alternative formulation of the underlying optimization problem would be necessary. However, in such a scenario, estimators have only been provided for the limiting cases of tail dependence: independence and comonotonicity. We extend the estimation of multivariate extreme expectiles (MEEs) by providing a consistent estimation scheme for random vectors with any arbitrary dependence structure. Specifically, we show that if the upper tail dependence function, tail index, and tail ratio can be consistently estimated, then one would be able to estimate MEEs accurately. The finite-sample performance of this methodology is illustrated using both simulated and real data.

E0226: **A bias-reduced version of the Weissman estimator for extreme value-at-risk**
Presenter: **Jonathan El Methni**, Universite de Paris, France
Co-authors: Stephane Girard

One of the most popular risk measures is the Value-at-Risk, which in statistical terms corresponds to the upper α-quantile of the distribution where $\alpha \in (0,1)$ is the risk level. We are interested in estimating this risk measure at an extreme level which means when $alpha$ tends to 0 as the sample size goes to infinity. If the random variable of interest has a heavy-tailed distribution, a common estimator of the Value-at-Risk at extreme levels is the Weissman estimator. The latter is based on two estimators: an order statistic to estimate an intermediate quantile and an estimator of the tail index. The usual practice is to select the same intermediate sequence for both estimators. We show how an adapted choice of two different intermediate sequences leads to a reduction of the asymptotic bias associated with the resulting Weissman estimator. Our approach is compared to other bias-reduced estimators of the extreme Value-at-Risk both on simulations and on a financial real dataset.

E0447: **Automatic threshold selection for extreme value regression models**
Presenter: **Julien Hambuckers**, University of Liege, Belgium
Co-authors: Antoine Usseglio-Carleve, Marie Kratz

The problem of threshold selection is investigated in the context of the extreme value regression model. In this regression context, the threshold choice is a non-trivial task since it should also depend on the covariates and can have important consequences on the final estimates. We propose an efficient and robust solution to automatically estimate these thresholds with the help of the distributional regression machinery. We illustrate its properties through several simulation studies. The method is later applied to the estimation of hedge funds tail risks.

E0914: **Heavy-tailed extremile regression in risky seismic areas**
Presenter: **Thibault Laurent**, Fondation Jean-Jacques Laffont, France
Co-authors: Abdelaati Daouia, Gilles Stupfler

Extremile regression defines a least-squares analogue of quantile regression as is the case in the duality between the conditional mean and the conditional median. The use of extremiles appears naturally in risk handling where they enjoy various intuitive meanings in terms of weighted moments rather than tail probabilities. They account for the magnitude of infrequent observations and not only for their relative frequency. They belong to both classes of concave distortion risk measures and coherent spectral risk measures of law–invariant type. We study the implications of extremile regression for estimating tail risk, focusing on heavy-tailed seismic distributions in risky areas. Based on a localized and averaged Hill estimator of the underlying conditional tail index, we present extrapolated estimators for high conditional extremiles and derive their asymptotic normality under mild conditions. In a simulation study, we examine their performance on finite samples in comparison with a method based on a kernel smoothing estimator of the conditional tail index. On an earthquake dataset in Indonesia and its surroundings in the Indian Ocean, these estimators guarantee a more reasonable and prudent differentiation of the severity of massive earthquakes geographically compared to the traditional Value at Risk and Tail Conditional Mean.

EO752 Room K2.31 Nash (Hybrid 07) STATISTICAL METHODS FOR MENDELIAN RANDOMIZATION Chair: Apostolos Gkatzionis

E0805: **Two-sample MR: Correction for winner's curse and weak instruments bias for unknown degree of sample overlap**
Presenter: **Ninon Mounier**, University of Lausanne, Switzerland
Co-authors: Zoltan Kutalik

Inverse-variance weighting (IVW) two-sample Mendelian Randomization (MR) is the most widely used method to estimate the causal effect of an exposure on an outcome. However, the resulting causal effect estimates may suffer from winners curse and weak instruments, and the extent of these biases is influenced by the degree of sample overlap, which is often unknown. Assuming a spike-and-slab genomic architecture, the bias of the IVW estimator can be analytically derived. This bias is driven by two forces: one acting towards the null independently of sample overlap and a second, proportional to the degree of overlap and the phenotypic correlation. We can estimate it using only summary statistics. Hence we propose a correction of the IVW-MR estimate and compare it against its uncorrected counterpart under a wide range of simulation settings. Finally, we applied our approach to 272 pairs of traits from UKBB. Using simulated data, we observed significant differences between IVW-MR and corrected effects, for all degrees of overlap. In all explored scenarios, our correction reduced the bias, sometimes even by up to 30 folds. When applied

to UKBB traits, we observed significant differences ($p < 0.05/272$) between IVW-MR and corrected effects for 15% of pairs. For example, we showed that the effect of educational attainment on BMI was 30% larger after correction ($\alpha_IVW = -0.462, \alpha_corrected = -0.602$).

E0525: **A Bayesian approach to overlapping-sample Mendelian randomization**
Presenter: **Hui Guo**, University of Manchester, United Kingdom

Mendelian randomization (MR) is a popular approach to causal inference in medical research. It uses exposure associated genetic variants, or most commonly, single nucleotide polymorphisms (SNPs) as instruments to investigate causal relationships between exposures and outcomes. Existing MR methods and their platforms have mainly focused on using publicly available summary statistics of the SNP-exposure association and SNP-outcome association obtained from two independent studies (namely two-sample MR). However, it has been brought to our attention that there are cases where a subgroup of participants was included in both of the studies, especially when data were collected at the population level. To enable a two-sample MR analysis, it is common practice that one study is discarded and a third study used instead of such that the studies are independent. Or, if data are available at the individual level, summary statistics are estimated from association analysis using data only from the non-overlapping participants. We will introduce a Bayesian MR approach that converts a two-sample or overlapping sample case into a one-sample setting, where the unmeasured data are treated as unknown parameters in the model that can be imputed in Markov chain Monte Carlo. It allows us to use all the available data without removing any participants or study. By its nature, Bayesian approach offers a more flexible way of dealing with complex models (e.g., multiple exposures, pleiotropy).

E0980: **Avoiding bias in Mendelian randomization when stratifying on a collider**
Presenter: **Claudia Coscia**, Spanish National Cancer Research Centre (CNIO), Spain
Co-authors: Dipender Gill, Teresa Perez, Nuria Malats, Stephen Burgess

Mendelian randomization (MR) uses genetic variants as instrumental variables to investigate the causal effect of a risk factor on an outcome. A collider is a variable influenced by two or more variables. A naive calculation of MR estimates in strata of the population defined by colliders may generate collider bias. We propose an approach that allows stratifying the MR analysis avoiding bias. The approach constructs a new variable, the residual-collider, as the residual from regressing the collider on the instrument, and calculates causal estimates in strata defined by quantiles of the residual-collider. Stratum-specific collider and residual-collider estimates will have equivalent interpretations, but residual-collider estimates will not suffer from collider bias. We apply this approach in several simulation scenarios considering different characteristics of the collider variable and instrument strengths, and to investigate the causal effect of smoking on bladder cancer in strata of the population defined by bodyweight. The approach generated unbiased estimates in all simulation settings, and a trend in the stratum-specific MR estimates at different bodyweight levels that suggested stronger effects of smoking on bladder cancer among individuals with lower bodyweight. This approach can be used to perform MR studying heterogeneity among subgroups of the population while avoiding collider bias.

E1340: **Sparse dimensionality reduction approaches in Mendelian randomization for highly correlated exposures**
Presenter: **Vasilis Karageorgiou**, University of Exeter, United Kingdom
Co-authors: Dipender Gill, Jack Bowden, Verena Zuber

Multivariable Mendelian randomization (MVMR) is an instrumental variable technique that generalizes the MR framework for multiple exposures. It is subject to the pitfall of multi-collinearity. The efficiency of MVMR estimates thus depends on the correlation of exposures. Dimensionality reduction techniques such as principal component analysis (PCA) provide transformations of the included variables such that they are effectively uncorrelated. We propose the use of sparse PCA (sPCA) algorithms for obtaining uncorrelated transformations of the exposures and can consequently provide more reliable summary-level estimates. The approach consists of three steps. We first apply a sparse dimensionality method and transform the SNP-exposure summary statistics into latent components. We choose a smaller number of latent components based on data-driven cutoffs, and estimate their strength as combined instruments with a novel F-statistic. Finally, we perform two-sample MR using the transformed exposures. This pipeline is demonstrated in a simulation study of highly correlated exposures and an applied example using summary data from a genome-wide association study on 118 highly correlated metabolites. As a positive control, we tested the causal effects of the transformed exposures on coronary heart disease. Compared to the conventional inverse-variance weighted MVMR method, sparse component analysis achieved a better balance of sparsity and a biologically insightful grouping of the traits.

E0903: **A latent mixture model for heterogeneous causal mechanisms in mendelian randomization**
Presenter: **Daniel Iong**, University of Michigan, Ann Arbor, United States
Co-authors: Qingyuan Zhao, Yang Chen

Mendelian Randomization (MR) is a popular method in epidemiology and genetics that uses genetic variation as instrumental variables for causal inference. Existing MR methods usually assume most genetic variants are valid instrumental variables that identify a common causal effect. There is a general lack of awareness that this effect homogeneity assumption can be violated when there are multiple causal pathways involved, even if all the instrumental variables are valid. We introduce a latent mixture model MR-PATH that groups instruments that yield similar causal effect estimates together. We develop a Monte-Carlo EM algorithm to fit this mixture model, derive approximate confidence intervals for uncertainty quantification, and adopt a modified Bayesian Information Criterion (BIC) for model selection. We verify the efficacy of the Monte-Carlo EM algorithm, confidence intervals, and model selection criterion using numerical simulations. We identify potential mechanistic heterogeneity when applying our method to estimate the effect of high-density lipoprotein cholesterol on coronary heart disease and the effect of adiposity on type II diabetes.

EO619 Room K2.40 (Hybrid 08) INTERFACE BETWEEN BAYESIAN STATISTICS AND MACHINE LEARNING Chair: Sara Wade

E0651: **Bayesian nonparametric methods for conditional independence testing**
Presenter: **Sarah Filippi**, Imperial College London, United Kingdom

Present Bayesian nonparametric methods for hypothesis testing are presented. In particular, we will focus on methods for quantifying the relative evidence in a dataset in favour of the dependence or independence of two variables conditionally on other variables. The approaches use Polya tree priors on spaces of probability densities, accounting for uncertainty in the form of the underlying distributions in a nonparametric way. The Bayesian perspective provides an inherently symmetric probability measure of conditional dependence or independence, a feature particularly advantageous in causal discovery.

E0791: **Identifiable variational autoencoders via sparse decoding**
Presenter: **Gemma Moran**, Columbia University, United States
Co-authors: Dhanya Sridhar, Yixin Wang, David Blei

Consider unsupervised representation learning: given datapoints of high-dimensional features, we want to learn low dimensional factors – a representation – that captures the observed data. We consider sparse representation learning, where each latent factor influences a subset of features. This notion of sparsity often reflects underlying patterns in data; in movie-ratings data, for example, each movie (feature) is only described by a few genres (factors). To this end, we introduce the Sparse Variational Autoencoder (Sparse VAE), a deep generative model with priors that encourage features to depend on only a few factors. The main technical result is proving that the Sparse VAE is identifiable: given data drawn from the model, there exists a unique optimal set of factors. This result sets the Sparse VAE apart from many deep generative models for representation learning, which are unidentifiable. One key assumption is the existence of "anchor features": for each factor, there exist features that depend only on that factor. Importantly, these anchor features do not need to be known a priori. We empirically study the Sparse VAE with simulated data and show

that it recovers the true latent factors when related methods do not. We study movie rating and text datasets, and show that the Sparse VAE predicts well on holdout data as well as data drawn from a different test distribution.

E0935: **Dimension-robust neural network priors for Bayesian reinforcement learning**
Presenter: **Torben Sell**, University of Edinburgh, United Kingdom
Co-authors: Sumeetpal Singh

A new neural network-based prior is discussed for real-valued functions on R^d which, by construction, is more easily and cheaply scaled up in the domain dimension d compared to the usual Karhunen-Loeve function space prior a property we refer to as "domain dimension robustness". The new prior is a Gaussian neural network prior, where each weight and bias has an independent Gaussian prior, but with the key difference that the variances decrease in the width of the network in such a way that the resulting function is almost surely well defined in the limit of an infinite-width network. We show that in a Bayesian treatment of inferring unknown functions, the induced posterior over functions is amenable to Monte Carlo sampling using Hilbert space Markov chain Monte Carlo (MCMC) methods. This type of MCMC is popular, e.g. in the Bayesian Inverse Problems literature, because it is stable under mesh refinement, i.e. the acceptance probability does not shrink to 0 as more parameters of the functions prior are introduced, even ad infinitum. We also implement examples in Bayesian Reinforcement Learning to automate tasks from data and demonstrate, for the first time, the stability of MCMC to mesh refinement for these types of problems.

E0945: **Variational Bayes for high-dimensional linear regression with sparse priors**
Presenter: **Kolyan Ray**, Imperial College London, United Kingdom
Co-authors: Botond Szabo

A core problem in Bayesian statistics is approximating difficult-to-compute posterior distributions. In variational Bayes (VB), a method from machine learning, one approximates the posterior through optimization, which is typically faster than Markov chain Monte Carlo. We study a mean-field (i.e. factorizable) VB approximation to Bayesian model selection priors, including the popular spike-and-slab prior, in sparse high-dimensional linear regression. We establish convergence rates for this VB approach, studying conditions under which it provides good estimation. We also discuss computational issues and study the empirical performance of the algorithm.

EO669 Room K2.41 (Hybrid 09) SHRINKAGE PRIORS FOR STRUCTURED VARIABLES (VIRTUAL) Chair: Mahlet Tadesse

E0999: **Nonparametric smoothing with Markov random fields and shrinkage priors**
Presenter: **James Faulkner**, National Oceanic and Atmospheric Administration, United States

A locally-adaptive nonparametric curve fitting method is presented that operates on similar principles as structured variable selection. This method uses shrinkage priors to induce sparsity in order-k differences in the latent trend function, providing a combination of local adaptation and global control. Using a scale-mixture representation of shrinkage priors, we represent our method as a form of kth-order Markov random field smoothing. This formulation offers computational advantages and allows the application of our method where Gaussian Markov random fields have previously been used. We use Hamiltonian Monte Carlo for posterior inference because it provides superior performance in the presence of the high dimensionality and strong parameter correlations exhibited by our models. We compare the performance of three prior formulations using simulated data and find the horseshoe prior provides the best compromise between bias and precision. We discuss a few extensions of the models, including an extension to the two-dimensional spatial setting.

E0698: **Graph-structured variable selection with Gaussian Markov random field horseshoe prior**
Presenter: **Marie Denis**, CIRAD, Georgetown University, France
Co-authors: Mahlet Tadesse

A graph structure is commonly used to characterize the dependence between variables. The Bayesian approach provides a natural framework to integrate the graph information through the prior distributions. We present an approach that combines Gaussian Markov random field (MRF) prior with global-local (GL) shrinkage prior for the selection of graph-structured variables. The local shrinkage parameters capture the dependence between connected covariates and take into account the sign of their empirical correlations. This encourages a similar amount of shrinkage for the regression coefficients while allowing them to have opposite signs. For non-connected variables, a standard horseshoe prior is specified. We illustrate the performance of the model with simulated data and real data applications, one in quantitative trait loci mapping with dependence between adjacent genetic markers and the other in gene expression data with a general estimated dependence structure between genes.

E0972: **The dynamic triple gamma prior**
Presenter: **Peter Knaus**, WU Vienna University of Economics and Business, Austria

Time-varying parameter (TVP) models are widely used in time series analysis for their ability to capture gradual changes in the effect of explanatory variables on an outcome variable of interest. The high degree of flexibility they offer can lead to overfitting when not properly regularized, which in turn results in poor out of sample predictive performance. On the other hand, approaches that are too restrictive risk not letting salient features of the data filter through. In light of these requirements, we propose a novel shrinkage process for sparse state space and TVP models. Building on previous work, we leverage the desirable properties of the triple gamma prior and introduce a shrinkage process that aims to combine sufficient regularization with enough flexibility to capture salient features of the data. Links to the literature are explored and an efficient MCMC algorithm is discussed.

E1175: **Structured shrinkage priors**
Presenter: **Maryclare Griffin**, University of Massachusetts Amherst, United States
Co-authors: Peter Hoff

In many regression settings, the unknown coefficients may have some known structure, for instance, they may be ordered in space or correspond to a vectorized matrix or tensor. At the same time, the unknown coefficients may be sparse, with many nearly or exactly equal to zero. However, many commonly used priors and corresponding penalties for coefficients do not encourage simultaneously structured and sparse estimates. We develop structured shrinkage priors that generalize multivariate normal, Laplace, exponential power and normal-gamma priors. These priors allow the regression coefficients to be correlated a priori without sacrificing elementwise sparsity or shrinkage. The primary challenges in working with these structured shrinkage priors are computational, as the corresponding penalties are intractable integrals and the full conditional distributions that are needed to approximate the posterior mode or simulate from the posterior distribution may be non-standard. We overcome these issues using a flexible elliptical slice sampling procedure, and demonstrate that these priors can be used to introduce structure while preserving sparsity.

EC857 Room K0.50 (Hybrid 06) CONTRIBUTIONS IN TIME SERIES Chair: Enea Bongiorno

E1379: **New approach to multistep ahead forecasting: What to do when neither direct nor recursive methods are fit to do the job**
Presenter: **Julija Tastu**, Maersk Line, Denmark

There are two classical ways of getting multistep ahead forecasts: direct and iterative. The direct method requires fitting a separate model for each of the prediction horizons of interest. The iterative one relies on describing system dynamics one step ahead and using predictions at the current step to determine the process value in the next step. Hybrid approaches exist and are based on either including iterative predictions as input to direct models or modelling all the horizons jointly as a multivariate process. None of the above methods is applicable when facing the following practical issue. Suppose making direct models is simply too costly computationally and suppose that iterative approach is not good enough as

error propagation caused by iterations becomes too high. In this case, none of the alternatives mentioned above offers a solution. We offer an alternative. It is based on fitting direct models to several horizons, fitting points, and assuming that model coefficients are smooth functions of the forecast horizon. Then predictions for horizons outside of the fitted ones can be obtained by interpolation of the results from the fitting points. This approach is a computationally lighter version of the direct method. On top of computational benefits, it offers a potential to improve forecast accuracy by smoothing out more noise from the signal than usual direct models offer.

E0266: **Robust inference for change points in piecewise polynomials of general degrees**
Presenter: **Shakeel Gavioli-Akilagun**, London School of Economics, United Kingdom
Co-authors: Piotr Fryzlewicz

Multiple change-point detection has become popular with the routine collection of complex non-stationary time series. An equally important but comparatively neglected question concerns quantifying the level of uncertainty around each putative changepoint. Though a handful of procedures exist in the literature, most all make assumptions on the density of the contaminating noise which are impossible to verify in practice. We present a procedure that, under minimal assumptions, returns localized regions of a data sequence that must contain a changepoint at some global significance level chosen by the user. Our procedure is computationally efficient, applicable to change points in higher-order polynomials, and moreover, all results are fully non-asymptotic. We will discuss some appealing theoretical properties of our procedure, and show its good practical performance on real and simulated data.

E1592: **Missing data imputation via state space model for non-stationary multi-variate time series in mHealth**
Presenter: **Linda Valeri**, Columbia University, United States
Co-authors: Xiaoxuan Cai

Missing data is a ubiquitous problem in biomedical and social science research. Data imputation is a commonly recommended remedy. Mobile technology (e.g., mobile phones and wearable devices) allows to closely monitoring individuals behavior and symptoms in real-time, and holds great potential for scientific discoveries and personalized treatment. Continuous data collection using mobile technology gives rise to a new type of data, entangled multivariate time series of outcome, exposure and covariates, and poses new challenges in missing data imputation for valid inference on treatment effects. Most existing imputation methods are either designed for longitudinal data with a limited number of follow-ups or for stationary time series, which may not be suitable in the field of psychiatry when mental health symptoms display dramatic changes over time or patients experience shifts in treatment regime over their course of recovery. We propose a novel imputation method based on the state-space model (SSMimpute) to tackle missingness in outcomes when multivariate time series are potentially non-stationary. We evaluate its theoretical properties and performance in extensive simulations, showing its advantages over other commonly used strategies for missing data. We apply the SSMimpute method in the analysis of a multi-year observational smartphone study of bipolar patients to evaluate the association between social network size and psychiatric symptoms adjusting for confounding.

CI008 Room K E. Safra (Multi-use 01) NEW DEVELOPMENTS IN HIGH-DIMENSIONAL ECONOMETRICS (HYBRID) Chair: Degui Li

C0168: **Estimating the lasso's effective noise**
Presenter: **Michael Vogt**, University of Ulm, Germany
Co-authors: Johannes Lederer

Much of the theory for the lasso in the high-dimensional linear model $Y = X\beta^* + \varepsilon$ hinges on the quantity $2\|X^\top\varepsilon\|_\infty/n$, which we call the lasso's effective noise. Among other things, the effective noise plays an important role in finite-sample bounds for the lasso, the calibration of the lasso's tuning parameter, and inference on the parameter vector β^*. We develop a bootstrap-based estimator of the quantiles of the effective noise. The estimator is fully data-driven; that is, it does not require any additional tuning parameters. We equip our estimator with finite-sample guarantees and apply it to tuning parameter calibration for the lasso and to high-dimensional inference on the parameter vector β^*.

C0167: **Factor-adjusted network analysis for high-dimensional time series**
Presenter: **Haeran Cho**, University of Bristol, United Kingdom
Co-authors: Matteo Barigozzi, Dom Owens

A methodology is proposed for modelling network structures of high-dimensional time series exhibiting strong serial- and cross-sectional correlations. We adopt a factor-adjusted vector autoregressive (VAR) model where, after the factors account for pervasive co-movements of the variables, remaining idiosyncratic dependence between the variables is modelled by as a sparse VAR process. We propose methods for estimating the latent VAR model and thus learning a directed network representing the Granger causal linkages between the variables, an undirected one embedding the contemporaneous relationships among the residuals and finally, one that summarises both lead-lag and contemporaneous linkages by means of the long-run partial correlations. We provide consistency with rates of all estimated quantities without any specific distributional assumption but by requiring only the existence of fourth moments. As a by-product, the first complete treatment of factor-adjusted sparse VAR estimation is offered. Simulations and real data applications are provided to demonstrate the good performance of the proposed methodology.

C1606: **Nonparametric estimation of large spot volatility matrices for high frequency data**
Presenter: **Ruijun Bu**, University of Liverpool, United Kingdom
Co-authors: Degui Li, Oliver Linton, Hanchao Wang

The focus is on estimating spot/instantaneous volatility matrices of high-frequency data collected for a large number of assets. We first combine classic nonparametric kernel-based smoothing with a generalised shrinkage technique in the matrix estimation for noise-free data under a uniform sparsity assumption, a natural extension of the approximate sparsity commonly used in the literature for low-frequency data. The uniform consistency property is derived for the proposed spot volatility matrix estimator with convergence rates comparable to the optimal minimax one. For the high-frequency data contaminated by microstructure noise, we introduce a localised pre-averaging estimation method in the high-dimensional setting which first pre-whitens data via a kernel filter and then uses the estimation tool developed in the noise-free scenario. Furthermore, we apply the developed technique to estimate the time-varying volatility matrix of the high-dimensional noise vector, and establish the relevant uniform consistency result. Numerical studies are provided to examine the performance of the proposed estimation methods in finite samples.

CO412 Room Virtual R31 SIGNAL EXTRACTION Chair: Anindya Roy

C0797: **A Bayesian framework for handling outliers in seasonal adjustment**
Presenter: **Anindya Roy**, University of Maryland Baltimore County, United States
Co-authors: Tucker McElroy

Current seasonal adjustment approaches require the identification of extreme values and outliers as fixed effects as an initial step, followed by their removal. The extreme value adjusted series are then filtered using linear techniques ideally suited for Gaussian observations. The final analysis however ignores the added uncertainty due to the specification of the time of occurrences of the outliers and also the effect of inference from the removal of the outliers. Alternatively, the outliers can be modeled as arising from latent stochastic processes driven by heavy-tailed innovations; extraction of latent components then follows non-linear techniques, and does not require identification of extreme epochs. We propose a Bayesian framework for modeling the dynamic components of a time series along with the extreme observations. The approach also yields a posterior for a counterfactual trend estimate that helps evaluate the impact of extreme occurrences.

C0821: **Zero-crossings of time series: Sign-prediction, mean-square error and a holding-time constraint**
Presenter: **Marc Wildi**, Zurich University, Switzerland

An extension of classic time series approaches is proposed which addresses zero-crossings of a zero-mean stationary time series. Specifically, we subject the original optimization criterion to a novel holding-time constraint which conditions the expected duration between consecutive crossings. Formally, the solution of this prediction problem is obtained by overlaying the classic estimate with a novel smoothing design whereupon the relative weight assigned to the latter by the criterion depends on the strength of the holding-time constraint. The resulting tradeoff is part of a general trilemma addressing smoothing, forecast accuracy and timeliness in prediction. Besides an analysis based on simple forecast and signal extraction exercises we also provide an application of our novel approach to business-cycle analysis. The examples illustrate that improved smoothing-capability (noise suppression) does not necessarily conflict with timeliness (relative lead) in this extended methodological framework.

C0967: **Enhanced methods of seasonal adjustment**
Presenter: **Stephen Pollock**, University of Leicester, United Kingdom

The effect of the conventional model-based methods of seasonal adjustment is to nullify the elements of the data that reside at the seasonal frequencies and to attenuate the elements at the adjacent frequencies. It may be desirable to nullify some of the adjacent elements instead of merely attenuating them. For this purpose, two alternative sets of procedures are presented that have been implemented in a computer program named SEASCAPE. In the first set of procedures, a basic seasonal adjustment filter is augmented by additional filters that are targeted at the adjacent frequencies. In the second set of procedures, a Fourier transform of the data is exploited to allow the elements in the vicinities of the seasonal frequencies to be eliminated or attenuated at will. The question is raised of whether an estimated trend-cycle trajectory that is devoid of high-frequency noise can serve in place of the seasonally adjusted data.

C1164: **A randomized missing data approach to robust filtering and forecasting**
Presenter: **Dobrislav Dobrev**, Federal Reserve Board, United States
Co-authors: Derek Hansen, Pawel Szerszen

A simple new randomized missing data (RMD) approach is put forward to robust filtering of state-space models, motivated by the idea that the inclusion of only a small fraction of available highly precise measurements can still extract most of the attainable efficiency gains for filtering latent states, estimating model parameters, and producing out-of-sample forecasts. In our general RMD framework we develop two alternative implementations: endogenous (RMD-N) and exogenous (RMD-X) randomization of missing data. A degree of robustness to outliers and model misspecification is achieved by purposely randomizing over the utilized subset of seemingly highly precise but possibly misspecified or outlier contaminated data measurements in their original time-series order, while treating the rest as if missing. Time-series dependence is thus fully preserved and all available measurements can get utilized subject to a degree of downweighting depending on the loss function of interest. The arising robustness-efficiency trade-off is controlled by varying the fraction of randomly utilized measurements or the incurred relative efficiency loss. As an empirical illustration, we show consistently attractive performance of our RMD framework in popular unobserved components models for extracting inflation trends. We further consider model extensions that more directly reflect inflation targeting by central banks and reveal its effectiveness through improved inflation forecasting.

CO280 Room Virtual R32 VOLATILITY COMPONENT MODELS Chair: Christian Conrad

C0249: **A reality check on the GARCH-MIDAS volatility models**
Presenter: **Nader Virk**, Plymouth Business School, United Kingdom

A battery of model evaluation tests is employed for a broad set of GARCH-MIDAS models and account for data snooping bias. We document that, inferences based on standard tests for GM variance components can be misleading. Our data mining free results show that the gains of macro-variables in forecasting total (long run) variance by GM models are overstated (understated). Estimation of different components of volatility is crucial for designing differentiated investing strategies, risk management plans and pricing of derivative securities. Therefore, researchers and practitioners should be wary of data-mining bias, which may contaminate a forecast that may appear statistically validated using robust evaluation tests.

C0284: **Hypotheses testing in mixed-frequency volatility models: A bootstrap approach**
Presenter: **Vincenzo Candila**, Sapienza University of Rome, Italy
Co-authors: Lea Petrella, Alberto Arcagni

It is widely recognized that standard likelihood-based inference suffers from the presence of nuisance parameters. This problem is particularly relevant in the context of Mixing-Data Sampling (MIDAS) methods applied to volatility modeling. In this framework, the volatility can be decomposed into two components, one varying daily and another varying according to the (lower) frequency of the additional volatility determinant. The MIDAS methods estimate the weights associated with each lagged realization of the low-frequency variable. A problem arises when the interest is on testing the whole impact of the low-frequency variable because, under the null hypothesis of no influence, the weight parameters are not identifiable. From this aspect, the weight parameters can be seen as nuisance parameters. This situation interferes with the asymptotic distribution of the common statistical tests employed to evaluate the significance of all the model's parameters. In order to overcome this problem, a bootstrap likelihood ratio (BLR) test is proposed, simulating the likelihood ratio test distribution. Using a Monte Carlo experiment, the proposed BLR test presents considerably good performances in terms of the test's size and power, generally better than the standard likelihood ratio test.

C0724: **Volatility estimation when the zero-process of financial return is nonstationary**
Presenter: **Genaro Sucarrat**, BI Norwegian Business School, Norway
Co-authors: Christian Francq

Financial returns are frequently nonstationary due to the nonstationary distribution of zeros. In daily stock returns, for example, the nonstationarity can be due to an upwards trend in liquidity over time, which may lead to a downwards trend in the zero-probability. In intraday returns, the zero-probability may be periodic: It is lower in periods where the opening hours of the main financial centres overlap, and higher otherwise. A nonstationary zero-process invalidates standard estimators of volatility models, since they rely on the assumption that returns are strictly stationary. We propose a GARCH model that accommodates a nonstationary zero-process, derive a 0-adjusted QMLE for the parameters of the model, and prove its consistency and asymptotic normality under mild assumptions. The volatility specification in our model can contain higher-order ARCH and GARCH terms, and past zero-indicators as covariates. Simulations verify the asymptotic properties in finite samples, and show that the standard estimator is biased. An empirical study of daily and intraday returns illustrate our results. They show how a nonstationary zero-process induces time-varying parameters in the conditional variance representation, and that the distribution of zero returns can have a strong impact on volatility predictions.

C1104: **Inference on multiplicative component GARCH without any small-order moment**
Presenter: **Baye Matar Kandji**, CREST/Institut Polytechnique de Paris, France
Co-authors: Christian Francq, Jean-Michel Zakoian

The aim is to investigate the existence of strictly stationary solutions and the asymptotic properties of Quasi-Maximum Likelihood (QML) estimation for a class of multiplicative two-component (short-term and long-run volatilities) GARCH models. We show that the strict stationarity condition is compatible with the infiniteness of any small-order power moment, contrary to the classical GARCH setting. The strong consistency

and asymptotic normality of the QML estimator are established under mild conditions. The results are illustrated via Monte Carlo experiments and real financial data.

CO150 Room Virtual R33 ADVANCES IN MACROECONOMETRICS Chair: Saeed Zaman

C0197: **Deep quantile regression**
Presenter: **Ilias Chronopoulos**, King's College London, United Kingdom
Co-authors: Aristeidis Raftapostolos, George Kapetanios

A deep quantile estimator is proposed using neural networks and their universal approximation property to examine a non-linear association between the conditional quantiles of a dependent variable and predictors. The proposed methodology is versatile and allows both the use of different penalty functions and high dimensional covariates. We present a Monte Carlo exercise where we examine the finite sample properties of the proposed estimator and show that our approach delivers good finite sample performance. We use the deep quantile estimator to forecast Value at Risk and find significant gains over linear quantile regression alternatives, supported by various testing schemes. We also contribute to the interpretability of neural networks output by making comparisons between the commonly used SHAP values and an alternative method based on partial derivatives.

C0277: **Decoupling shrinkage and selection for the Bayesian quantile regression**
Presenter: **Tibor Szendrei**, Heriot-Watt University, United Kingdom
Co-authors: David Kohns

The idea of decoupling shrinkage and sparsity for continuous priors is extended to Bayesian Quantile Regression (BQR). The procedure follows two steps: In the first step, we shrink the quantile regression posterior through state-of-the-art continuous priors and in the second step, we sparsify the posterior through an efficient variant of the adaptive lasso, the signal adaptive variable selection (SAVS) algorithm. We propose a new variant of the SAVS which automates the choice of penalisation through quantile specific loss-functions that are valid in high dimensions. We show in large scale simulations that our selection procedure decreases bias irrespective of the true underlying degree of sparsity in the data, compared to the un-sparsified regression posterior. We apply our two-step approach to a high dimensional growth-at-risk (GaR) exercise. The prediction accuracy of the un-sparsified posterior is retained while yielding interpretable quantile specific variable selection results. Our procedure can be used to communicate to policymakers which variables drive downside risk to the macroeconomy

C0450: **Macroeconomic forecasting with large stochastic volatility in mean VARs**
Presenter: **Chenghan Hou**, Hunan University, China

Vector autoregressions with stochastic volatility in both the conditional mean and variance are commonly used to estimate the macroeconomic effects of uncertainty shocks. Despite their popularity, intensive computational demands when estimating such models have made out-of-sample forecasting exercises impractical, particularly when working with large data sets. We propose an efficient Markov chain Monte Carlo (MCMC) algorithm for posterior and predictive inference in such models that facilitates such exercises. The key insight underlying the algorithm is that the (log-)conditional densities of the log-volatilities possess Hessian matrices that are banded. This enables us to build upon recent advances in band and sparse matrix algorithms for state-space models. In a simulation, we evaluate the new algorithm numerically and establish its computational and statistical efficiency over a conventional particle filter-based algorithm. Using macroeconomic data for the US, we find that such models generally deliver more accurate point and density forecasts over a conventional benchmark in which stochastic volatility only enters the variance of the model.

C0322: **A unified framework to estimate macroeconomic stars**
Presenter: **Saeed Zaman**, Federal Reserve Bank of Cleveland and University of Strathclyde, United States

A semi-structural time series model is developed to estimate jointly several macroeconomic "stars," i.e., unobserved long-run equilibrium levels of output (and growth rate of output), unemployment rate, the real rate of interest, productivity growth, price inflation, and wage inflation. The ingredients of the model are in part motivated by economic theory and, in part, by the empirical features necessitated due to the changing economic environment. We explicitly model the links between long-run survey expectations and stars to improve the stars econometric estimation. The approach permits time-variation in the relationships between various components, including time-variation in error variances. Our approach's by-products are the time-varying estimates of the wage and price Phillips curves, passthrough between prices and wages, which provide new insights into these empirical relationships' instability in US data. Generally, the contours of our stars echo those documented elsewhere in the literature – estimated using smaller models – but at times, the estimates of stars are different, and these differences have policy implications. Furthermore, the estimates of the stars are among the most precise. Lastly, we document the competitive forecasting properties of our model and, separately, the usefulness of stars estimates if they were used as steady-state values in external models.

CO726 Room Virtual R35 RECENT ADVANCES IN FINANCIAL ECONOMETRICS Chair: Seok Young Hong

C0373: **Efficient estimation of pricing kernels and market-implied densities**
Presenter: **Jeroen Dalderop**, University of Notre Dame, United States

The aim is to study the nonparametric identification and estimation of projected pricing kernels implicit in European option prices and underlying asset returns using conditional moment restrictions. The proposed series estimator avoids computing ratios of estimated risk-neutral and physical densities. Instead, we consider efficient estimation based on the conditional Euclidean empirical likelihood or continuously-updated GMM criterion, which takes into account the informativeness of option prices of varying strike prices beyond observed conditioning variables. In a second step, we convert the implied probabilities into predictive densities by matching the informative part of cross-sections of option prices. Empirically, pricing kernels tend to be U-shaped in the S&P 500 index return given high levels of the VIX, and call and ATM options are more informative about their payoff than put and OTM options.

C0377: **Augment large covariance matrix estimation with auxiliary network information**
Presenter: **Shaoran Li**, Peking University, China
Co-authors: Oliver Linton, Shuyi Ge, Weiguang Liu

The aim is to incorporate auxiliary information about the location of significant correlations into the estimation of 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 are now becoming available to researchers. Our Network Guided Estimator combines the banding and thresholding procedures with the help of augment information from other sources. Simulation results show that the new method has smaller estimation errors comparing with other methods in the literature. We empirically apply the Network Guided Estimator to estimate the covariance of the excess returns of S&P 500 stocks. The constructed global minimum variance portfolio has the smallest volatility among all competing methods.

C0439: **Estimation of nonstationary nonparametric regression model with multiplicative structure**
Presenter: **Ekaterina Smetanina**, University of Chicago, United States

A multiplicative nonstationary nonparametric regression model is presented, which allows for a broad class of nonstationary processes. We propose a three-step estimation procedure to uncover the conditional mean function and establish uniform convergence rates and asymptotic normality of our estimators. The new model can also be seen as a dimension reduction technique for a general two-dimensional time-varying nonparametric regression model, which is especially useful in small samples and for estimating explicitly multiplicative structural models. We consider two

applications: estimating a pricing equation for the US aggregate economy to model consumption growth, and estimating the shape of the monthly risk premium for S&P 500 Index data.

C0843: **Threshold regression for large datasets with common stochastic trends**
Presenter: **Daniele Massacci**, Kings College London, United Kingdom
Co-authors: Lorenzo Trapani

Inference is studied for threshold regression in the context of a large panel factor model with common stochastic trends. We develop a Least Squares estimator for the threshold level, deriving almost sure rates of convergence and proposing a novel, testing-based, way of constructing confidence intervals. We also investigate the properties of the PC estimator for the loadings and common factors in both regimes, and develop a procedure to estimate the number of common trends in each regime. Although the main focus is on common stochastic trends with mean zero, we show that our technique can be applied even in the presence of common factors with drifts and/or trend stationary common factors. The theoretical findings are corroborated through a comprehensive set of Monte Carlo experiments. Finally, the analysis of long-run returns from a set of buy and hold strategies shows the usefulness of our model for empirical work.

CO569 Room Virtual R38 CAUSAL AND NONCAUSAL TIME SERIES MODELS Chair: Alain Hecq

C0188: **Evaluation of the credibility of the Brazilian inflation targeting system using mixed causal-noncausal models**
Presenter: **Elisa Voisin**, Maastricht University, Netherlands
Co-authors: Alain Hecq, Joao Victor Issler

Predictive densities obtained via mixed causal-noncausal autoregressive models are used to evaluate the statistical sustainability of Brazilian inflation targeting system with the tolerance bounds. The probabilities give an indication of the credibility of the targeting system without requiring modelling people's beliefs. We also investigate the added value of including experts predictions of key macroeconomic variables.

C0203: **Inference in mixed causal and noncausal models withgeneralized Student's t-distributions**
Presenter: **Francesco Giancaterini**, Maastricht University, Italy
Co-authors: Alain Hecq

The aim is to analyze the properties of the Maximum Likelihood Estimator for mixed causal and noncausal models when the error term follows a Student's t-distribution. In particular, we compare several existing methods to compute the expected Fisher information matrix and show that they cannot be applied in the heavy-tail framework. For this purpose, we propose a new approach to make inferences on causal and noncausal parameters in nite sample sizes. It is based on the empirical variance computed on the generalized Student's t, even when the population variance is not nite. Monte Carlo simulations show the good performances of our new estimator for fat tail series. We illustrate how the different approaches lead to different standard errors in four-time series: annual debt to GDP for Canada, the variation of daily Covid-19 deaths in Belgium, the monthly wheat prices, and the monthly inflation rate in Brazil.

C0597: **Generalized covariance estimator**
Presenter: **Joann Jasiak**, York University, Canada
Co-authors: Christian Gourieroux

A class of semi-parametric dynamic models is considered with strong white noise errors including the standard Vector Autoregressive (VAR) model, the nonfundamental structural VAR model, the mixed causal-noncausal models and nonlinear dynamic models such as the (multivariate) ARCH-M model. For this class of models, we propose the Generalized Covariance (GCov) estimator, which is obtained by minimizing a residual-based multivariate portmanteau statistic. The GCov is a reliable alternative to the Generalized Method of Moments (GMM) providing semi-parametrically efficient estimates in one step. We derive the asymptotic properties of the GCov estimator and show that the associated residual-based portmanteau statistic is asymptotically chi-square distributed. The finite sample performance of the GCov estimator is illustrated in a simulation study. The estimator is also applied to a dynamic model of cryptocurrency prices.

C1159: **Is global warming (time) reversible?**
Presenter: **Alain Hecq**, Maastricht University, Netherlands
Co-authors: Francesco Giancaterini, Claudio Morana

Two new tests are proposed for time reversibility, exploiting the properties of mixed causal and noncausal models. Monte Carlo experiments show that they perform accurately in small samples, exhibiting a strong ability to detect time reversibility. Furthermore, the aim is to investigate whether time reversibility is a characteristic feature of the process of global warming. Using the aforementioned tests, time reversibility is tested on different climate time series: stratospheric aerosols from volcanic activity, global temperatures, temperature anomalies.

CO290 Room Virtual R40 TIME SERIES ECONOMETRICS Chair: Antonio Montanes

C0677: **Local warming: The globe vs Spain**
Presenter: **Lola Gadea**, University of Zaragoza, Spain
Co-authors: Jesus Gonzalo

Climate change and one of its main consequences, global warming, is a phenomenon that affects the whole planet, but the regional effects can be dramatic and require specific policies. Therefore, the analysis of Local warming is a fundamental complementary aspect of global warming which needs to be taken into account in the design of effective policies. The objective is to analyze Local Warming in the southernmost region of Europe which, according to all forecasts, will suffer the most adverse effects of climate change. Applying a previous methodology for studying the presence of increasing trends in the entire distribution of the temperature and not only in the mean, we also identify different types of warming. The empirical results obtained using different samples and time periods show, although with geographical nuances, a very clear warming and acceleration effect with increasing trends in all quantiles of the temperature distribution and an increase in dispersion due to the greater intensity of warming in the right upper tail of the distribution. This result differs from those found for the planet as a whole, and especially in the northern part of the Northern Hemisphere (the Arctic, the UK, etc). The type of global warming found previously can have unpredictable global consequences, but the type found here can have devastating local effects. For this reason, we need to think globally, and to act locally.

C0830: **Current account determinants in a globalized world**
Presenter: **Josep Lluis Carrion-i-Silvestre**, Universitat de Barcelona, Spain
Co-authors: Cecilio Tamarit, Mariam Camarero

Fresh evidence is presented on the long-term determinants of the current account overcoming some of the econometric flaws found in the previous empirical analysis as well as including the increasing importance of the financial integration for the adjustment of the current account. The intertemporal approach to the current account provides the underlying theoretical framework for this study. The sample contains eleven countries of the eurozone, ten developed non-European Union countries, and, finally, three emerging countries over the period 1972-2017. One distinctive feature of our single-equation-based analysis is the inclusion of foreign current accounts among the regressors as a way to capture the potential influence of the evolution of macroeconomic imbalances of other developed economies into the economy of interest. Using this information, it is possible to test the null hypothesis of weak cross-section dependence. The results show the distinctive importance of the usual variables with an increasing role of the dependence among countries in a globalized world.

C1424: **Causality between waste, recycling and GDP in G-7 countries: A bootstrap Granger non-causality test approach**

Presenter: **Alejandro Alcay**, University of Zaragoza, Spain

Co-authors: Antonio Montanes, Blanca Simon-Fernandez

The aim is to contribute to the literature on recycling-waste generation-economic growth causality. Previous results are characterized by their variability, particularly, across sample periods, sample sizes, and model specification. In order to overcome these issues, the causal links between these three variables for the G-7 countries are analyzed using bootstrap Granger non-causality tests with fixed size rolling subsamples. The data used includes annual total urban waste generation, recycling and real Gross Domestic Product (GDP) series. The results prove the importance of accounting for the presence of parameter instability, mainly due to the effect of the Great Recession. Our results encompass previous findings and offer an explanation for varying findings.

C1494: **Evaluation of classifiers through bilateral projections of the ROC curve**

Presenter: **Andres Romeu**, Universidad de Murcia, Spain

Co-authors: Maximo Camacho, Salvador Ramallo

Being a popular measure in other disciplines, the area under the ROC curve (AUC) has been recently proposed as a method to evaluate the classification ability of different indices in the business cycle literature. We show that the standard formulation of this measure poses some conceptual problems when there are anomalies such as either very large or very small differences in the scale of the signal along the time span. We propose an alternative measure based on adding a new dimension to the ROC, the threshold levels themselves. In this context, the AUROC is simply the projection on the true positive/false positive rate plane. Analogously, we consider the areas under the projections of the curve on the true positive vs. threshold plane and the false positive vs. threshold plane. We show that the difference between these two gives a measure of classifiers that is robust and able to discriminate abnormal signals that make little economic sense and that the standard AUROC fails to detect.

Sunday 19.12.2021 16:10 - 17:25 Parallel Session J – CFE-CMStatistics

EO838 Room K0.16 (Hybrid 02) STATISTICAL METHODS FOR ENVIRONMENTAL HEALTH DATA Chair: Carolina Euan

E0410: **Changepoint detection for home activity count data: A functional approach**
Presenter: **Israel Martinez-Hernandez**, Lancaster University, United Kingdom

Many companies aim to improve health and care of people by taking advantage of new technologies. We use a dataset from Howz; a company that helps older people to improve their health by identifying changes in their behavior over time. We propose a novel methodology to detect change points in the daily activities of people. Our goal is to detect changes across periods (days), and changes within each period are considered normal changes in daily activities. To this end, we model the cumulative activities for each period with a Cox process. Then, the sequence of the stochastic intensity functions results in a functional time series. Thus, our proposal uses this functional time series to detect changepoints across periods.

E1062: **Spatial disaggregation of disease count data**
Presenter: **Craig Anderson**, University of Glasgow, United Kingdom

Disease mapping focuses on estimating the spatial pattern of disease risk across a geographical region that has been subdivided into a set of administrative districts. The disease data consists of aggregated disease counts at this district level and traditionally this means that inference is also restricted to this geographical level. This standard inference can be susceptible to the modifiable areal unit problem (MAUP), whereby the estimated risk surface is affected by the arbitrary choice of district boundaries. The district-level count is really an aggregation of point level disease cases, and therefore if a different spatial partition of the region was selected we could observe different results. the aim is to address this problem by outlining a method for producing disaggregated disease risk estimates based on a regular 1km x 1km grid. The method is illustrated using an application to respiratory hospital admissions in Glasgow, Scotland.

E1073: **Optimal estimation of the sparsity index in Poisson size-biased sampling**
Presenter: **Laura Bondi**, University of Cambridge and Bocconi University, United Kingdom
Co-authors: Marco Bonetti, Marcello Pagano

If the probability that an individual is in the sample is proportional to a size variable, then we have size-bias. For example, when studying cancer history via a Cancer Registry, then a large family will have a higher chance of representation in the registry. Another example of this bias is provided by a study of the plague that hit Europe in 1630. When modeling the internment process into a plague ward (lazzaretto), an infected member of a household results in the whole household being admitted to the ward. The sparsity index (reciprocal of the intensity) is a parameter of interest. With size biased sampling caution must be taken when choosing an estimator. We explore these two examples and compute the uniformly minimum variance unbiased estimator for the sparsity index in size-biased Poisson sampling. We propose two exact algorithms, which are computationally burdensome even for small sample sizes. As an alternative, a third, approximate algorithm based on the inverse fast Fourier transform, is presented. An exact confidence interval based on the optimal estimator is also proposed. The performance of the estimation procedure is compared to classical maximum likelihood inference, both in terms of mean squared error and average coverage and width of the corresponding confidence intervals.

EO611 Room K0.19 (Hybrid 04) SINGLE-CELL RESOLUTION IMAGE ANALYSIS Chair: Simon Vandekar

E0798: **Computational tools to quantify and correct slide-to-slide variation in multiplexed immunofluorescence images**
Presenter: **Coleman Harris**, Vanderbilt University Medical Center, United States
Co-authors: Eliot McKinley, Joseph Roland, Qi Liu, Martha Shrubsole, Ken Lau, Robert Coffey, Julia Wrobel, Simon Vandekar

The multiplexed imaging domain is a nascent single-cell analysis field with a complex data structure susceptible to technical variability that disrupts inference. These in situ methods are valuable in understanding cell-cell interactions, but few computational tools exist to quantify or correct for technical variation in multiplexed imaging data. We implement and compare normalization algorithms in multiplexed imaging data, and present and evaluate the methods with a new R package, MxNorm. Our methods adapt the ComBat and functional data registration methods to remove slide effects in this domain, and we include a robust evaluation framework to compare the proposed approaches. The methods illustrate clear slide-to-slide variation in the raw, unadjusted data, demonstrating that many of the proposed normalization methods reduce this variation while preserving and improving the biological signal. The R package introduced here provides a clear framework to normalize and explore multiplexed imaging data, with methods to compare normalization algorithms and visualization tools to identify technical variation. Further, this framework can be integrated with any proposed normalization method or thresholding algorithm, providing open-source software to robustly improve data quality and evaluation criteria in the multiplexed domain.

E0907: **Quantification of spatial interaction between cancer and immune cells**
Presenter: **Inna Chervoneva**, Thomas Jefferson University, United States

Advanced analysis systems for pathology allow capturing spatial coordinates of all cells in immunohistochemistry images of the tumor microenvironment, but there are no established methods for objective quantification of spatial interaction between cancer and immune cells. The motivation comes from studies of cancer biomarkers in a tissue microarray of surgical specimens from a large cohort of ER+ breast cancer patients. We developed an objective thresholding algorithm for immune cell type classification and novel metrics of spatial interactions between cancer and immune cells based on distributions of the nearest neighbor distances. The spatial localization of CD8+ T cells and cancer cells was used to generate distributions of the nearest neighbor distances (NND) between cancer and CD8+ cells. The summary statistics of these distributions were considered as predictors of progression-free survival (PFS). The larger quantiles of NND were associated with shorter PFS, which is consistent with published results based on manual counts of tumor-infiltrating CD8+ lymphocytes. The median NND from cancer to CD8+ cells remained a significant predictor of PFS in the multivariable Cox model adjusted for known clinicopathological prognostic factors. The distributions of NND between cancer and various immune cells have the potential to provide novel cancer biomarkers.

E1096: **Bioconductor infrastructure for analyzing multiplex single cell imaging data in R**
Presenter: **Julia Wrobel**, Colorado School of Public Health, United States
Co-authors: Simon Vandekar, Coleman Harris

A new software ecosystem in R is introduced for data storage and spatial analysis of multiplex single-cell tissue imaging data. Our central R package is called spatialMI and builds on SpatialExperiment, which is an R package and S4 class on Bioconductor that provides a special data infrastructure for spatially resolved transcriptomics data that facilitates data storage, retrieval, subsetting, and interfacing with downstream tools. The spatialMI package adapts data structures from the SpatialExperiment class and inherits methods from that and the popular SingleCellExperiment class so that spatialMI users can easily access software developed for other similar single cell data types. For multiplex single-cell imaging data specifically, we build additional S4 methods to convert multichannel tiff images to a novel spatialMI data class so data from any multiplex platform, including CODEX, Vectra-Polaris, MIBI, IMC, etc, can be converted into the same type of Bioconductor data object. This facilitates consistency and ease of use in downstream analysis. We discuss the available functionality as well as forthcoming extensions in the form of satellite packages.

EO098 Room K0.20 (Hybrid 05) STATISTICAL METHODS FOR HIGH DIMENSIONAL NEUROIMAGING DATA II Chair: John Kornak

E0948: **Club Exco: Clustering brain extreme communities from multi-channel EEG data**
Presenter: **Matheus Guerrero**, King Abdullah University of Science and Technology, Saudi Arabia

Co-authors: Raphael Huser, Hernando Ombao

Epilepsy is a chronic neurological disorder affecting more than 50 million people globally. An epileptic seizure acts like a temporary shock to the neuronal system, disrupting regular electrical activity in the brain. Epilepsy is frequently diagnosed with electroencephalograms (EEGs). Current cluster methods for characterizing brain connectivity (e.g., coherence and partial coherence) are most influenced by the bulk of the EEG distribution rather than the tails. We develop the Club Exco method, which uses a spherical k-means procedure applied to the pseudo-angles derived from extreme amplitudes of EEG signals during an epileptic seizure to cluster brain extreme communities from multi-channel EEG data. With this approach, cluster centers can be interpreted as extremal prototypes, revealing the extremal dependence structures of communities of EEG channels. The clustering of channels can then be used as an exploratory tool to classify EEG channels into mutually asymptotically independent or asymptotically dependent groups. We apply the Club Exco method to investigate the differences in EEG brain connectivity networks between the pre- and post-seizure phases.

E1148: **Multivariate temporal point process regression**

Presenter: **Xiwei Tang**, University of Virginia, United States

Co-authors: Lexin Li

Point process modeling is gaining increasing attention, as point process type data are emerging in a large variety of scientific applications. Motivated by a neuronal spike trains study, we propose a novel point process regression model, where both the response and the predictor can be a high-dimensional point process. We model the predictor effects through the conditional intensities using a set of basis transferring functions in a convolutional fashion. We organize the corresponding transferring coefficients in the form of a three-way tensor, then impose the low-rank, sparsity, and subgroup structures on this coefficient tensor. These structures help reduce dimensionality, integrate information across different individual processes, and facilitate interpretation. We develop a highly scalable optimization algorithm for parameter estimation. We derive the large sample error bound for the recovered coefficient tensor, and establish the subgroup identification consistency, while allowing the dimension of the multivariate point process to diverge. We demonstrate the efficacy of our method through both simulations and a cross-area neuronal spike trains analysis in a sensory cortex study.

E0223: **Similarity-based multimodal regression**

Presenter: **Andrew Chen**, University of Pennsylvania, United States

Co-authors: Russell Shinohara, Haochang Shou

To better understand complex human phenotypes, large-scale studies have increasingly collected multiple data modalities across domains such as imaging, genomics, and physical activity. The properties of each data type often differ substantially and necessitate either multiple separate analyses or extensive processing to obtain comparable features for a single analysis. For a single data modality, multivariate distance matrix regression provides a distance-based framework for regression involving a wide range of data types. However, no distance-based method exists to handle multiple types of data in a single analysis. We extend a distance-based regression model to propose similarity-based multimodal regression (SiMMR), which enables simultaneous regression of multiple modalities through their distance profiles. We demonstrate through both simulation, imaging studies, and longitudinal wearable device analyses that our proposed method can detect associations in multimodal data of differing properties and dimensionalities, even with modest sample sizes. We furthermore perform experiments to evaluate several different test statistics and provide recommendations for applying our method in a wide range of scenarios.

EO080 Room Virtual R20 ADVANCES IN INFECTIOUS DISEASE MODELLING Chair: Rob Deardon

E0493: **Incorporating infectious duration-dependent transmission into Bayesian epidemic models**

Presenter: **Caitlin Ward**, University of Calgary, Canada

Compartmental models are commonly used to describe the spread of infectious diseases by estimating the probabilities of transitions between important disease states. A significant challenge in fitting Bayesian compartmental models lies in the need to estimate the duration of the infectious period, based on limited data providing only symptom onset date or another proxy for the start of infectiousness. Commonly, the exponential distribution is used to describe the infectious duration, an overly simplistic approach that is not biologically plausible. More flexible distributions can be used, but parameter identifiability and computational cost can worsen for moderately-sized or large epidemics. We present a novel approach that considers a curve of transmissibility over a fixed infectious duration. Incorporating infectious duration-dependent (IDD) transmissibility, which decays to zero over the infectious period, is biologically reasonable for many viral infections. Fixing the maximum length of the infectious period eases computational complexity in model fitting. Through simulation, we evaluate different functional forms of IDD transmissibility curves and show that the proposed approach offers an improved estimation of the time-varying reproductive number. The benefit of our approach is illustrated through a new analysis of the 1995 outbreak of Ebola Virus Disease in the Democratic Republic of the Congo.

E0696: **Inference for individual-level models of infectious diseases with an application to the COVID-19 pandemic**

Presenter: **Leila Amiri**, University of Manitoba, Canada

When individual-level data are available, more complex types of models may also be applied, then is otherwise the case. We initially formulate this in the context of the progression of an epidemic, for a disease where there is removal and no re-infection (i.e., a susceptible-exposed-infected-removed (SEIR) framework), as it is an appropriate model for the COVID-19. In the SEIR model, it is assumed that the infection rate will stay constant over time. We propose an individual-level statistical model (ILM) to rigorously predict the infection rate at each given time by incorporating the corresponding covariates and characteristics of infected people at the individual and area levels. We will then incorporate the infection rates at each given time into the SEIR model to accurately predict the outbreak over time that help policymakers for possible interventions. The performance of the proposed approach is evaluated through simulations. We also apply our proposed model to Manitoba COVID-19 datasets.

E1188: **Time series approaches to compare Covid-19 mortality in the province of Ontario, Canada, across three epidemic waves**

Presenter: **Charmaine Dean**, University of Waterloo, Canada

Co-authors: Georges Bucyibaruta, Dexen Xi, Elizabeth Renouf

The province of Ontario, Canada has experienced three epidemic waves of Covid-19, with a fourth wave currently underway. Hospitalizations and deaths, though lagging indicators, are considered an important metric for the comparison of waves. For the purposes of understanding the trend in hospitalizations and mortality, we utilize time series approaches for modeling outcomes. Cointegration analysis is employed to identify the long-run relationship between these processes. The outcomes are modeled through a shared latent stochastic error term in a novel framework that allows us to study the underlying correlation between two-time series processes. We also develop a logistic growth model for the cumulative number of deaths through each wave. Although an empirical model, it incorporates conceptual elements that support the framework required for modeling any infectious disease where hospitalizations are required in the management of the disease. By nature, the logistic growth model is deterministic, and so we induce stochasticity by incorporating the variability that is observed in modeling the daily counts and propose a tool that can be used to quantify the behavior of the disease within a short time period.

EO064 Room Virtual R21 RECENT DEVELOPMENT IN EXPERIMENTAL DESIGNS Chair: Po Yang

E1727: **Optimal experimental designs for efficiency of parameters and model discrimination**

Presenter: **Saumen Mandal**, University of Manitoba, Canada

In optimal design, we generally assume that the statistical model is known at the design stage. However, in many situations, this is not the case. We

need to implement a design that is efficient for two or more models, to discriminate between them, and select the best model. We consider different approaches to evaluate a criterion or a mixture of various criteria, and determine the efficiencies for the models and the parameters. We then consider a novel approach in which we optimize one objective subject to achieving a given efficiency for a parameter. We solve this constrained optimization problem by a Lagrangian approach. We eventually transform the problem to that of maximizing a number of functions simultaneously. These functions have a common maximum that is simultaneously attained at the optimum. We apply the method to some polynomial models, and discuss the results.

E1547: **Orthogonal array composite designs for drug combination experiments with applications for tuberculosis**
Presenter: **Jessica Jaynes**, California State University Fullerton, United States

The aim is to provide an overview of the orthogonal array composite design (OACD) methodology, show that they can be robust to missing data under practical scenarios and provide an application for tuberculosis. We compare the efficiencies of OACDs to the commonly used central composite designs (CCD) when there are a few missing observations and demonstrate that OACDs are more robust than the popular CCDs to missing observations for two scenarios. The first scenario assumes one observation is missing either from one factorial point or one additional point. The second scenario assumes two observations are missing either from two factorial points or from two additional points, or from one factorial point and one additional point. Two real-world applications of OACDs pertaining to tuberculosis are provided: a 155-run OACD with nine drugs and a 50-run OACD with six drugs.

E1502: **Structure of nonregular two-level designs**
Presenter: **David Edwards**, Virginia Commonwealth University, United States

Two-level fractional factorial designs are often used in screening scenarios to identify active factors. The block-diagonal structure of the information matrix of nonregular two-level designs is investigated. This structure is appealing since estimates of parameters belonging to different diagonal submatrices are uncorrelated. As such, the covariance matrix of the least-squares estimates is simplified and the number of linear dependencies is reduced. We connect the block diagonal information matrix to the parallel flats design literature and gain insights into the structure of what is estimable and/or aliased using the concept of minimal dependent sets. We show how to determine the number of parallel flats for any given design, and how to construct a design with a specified number of parallel flats using a Kronecker product construction. The usefulness of our construction method is illustrated by producing designs for the estimation of the two-factor interaction model with three or more parallel flats. We also provide a fuller understanding of recently proposed group orthogonal supersaturated designs. The benefits of parallel flats designs for analysis, including bias containment, are also discussed.

EO334 Room Virtual R22 MODEL ASSESSMENT II Chair: Maria Dolores Jimenez-Gamero

E0233: **A penalized multicollinearity measure for improved model assessment**
Presenter: **Kimon Ntotsis**, University of the Aegean, Greece
Co-authors: Alexandros Karagrigoriou, Andreas Artemiou

When it comes to variable interpretation, multicollinearity is among the biggest issues that must be surmounted, especially in this new era of Big Data Analytics. Since even moderate size multicollinearity can prevent a proper interpretation, special diagnostics must be recommended and implemented for identification purposes. Nonetheless, in the area of Finance and International Business, among other fields, these diagnostics are controversial concerning their "successfulness". It has been remarked that they frequently fail to do proper model assessment due to information complexity, resulting in model misspecification. The aim is to propose and investigate a robust and easily interpretable methodology, termed Elastic Information Criterion, capable of capturing multicollinearity rather accurately and effectively and thus providing a proper model assessment. The performance is investigated via simulated and real data.

E0278: **Spatialised similarity analysis of two digital elevation models via a large number of two-sample multinomial problems**
Presenter: **Virtudes Alba-Fernandez**, University of Jaen, Spain
Co-authors: Francisco Javier Ariza-Lopez, Maria Dolores Jimenez-Gamero

A Digital Elevation Model (DEM) is a bare earth elevation model representing the surface of the Earth. From the elevation model, other models can be derived, such as slopes, orientations, insolation, drainage networks or visual and watershed analysis. Usually, the comparison of two DEMs covering the same area is made by analysing the discrepancy between the two elevation models; however, we consider that a DEM is more than just elevations and that it also has spatialisation. Therefore, to make the comparison between two DEMs, a multivariate and spatialized perspective is proposed. Both the considered variables (categorized) and the way of spatialisation (natural or artificial) can be decided by the user. Motivated by this open question, the problem of testing a large number, say k, of multinomial two-sample problems is tackled. It will be assumed that the available data consist of k pairs of independent samples, which may be bounded or increase with k, but at a lower rate than k. A test statistic based on the Euclidean distance between the estimated proportions in each pair of populations is designed. The null distribution of the test statistic is unknown, and its asymptotic null distribution is derived. The asymptotic power is also studied. A simulation study is carried out to evaluate the behaviour of the proposal in a wide range of scenarios. Finally, a practical application is introduced using two DEM datasets from the same zone (Allo, Navarra, Spain).

E1605: **Post-selection inference for linear mixed model parameters using the conditional Akaike information criterion**
Presenter: **Katarzyna Reluga**, University of Toronto, Canada
Co-authors: Gerda Claeskens, Stefan Sperlich

The issue of post-selection inference for a fixed and a mixed parameter in a linear mixed model is investigated using a conditional Akaike information criterion as a model selection procedure. Within the framework of linear mixed models, we develop a complete theory to construct confidence intervals for regression and mixed parameters under three frameworks: nested and general model sets as well as misspecified models. The theoretical analysis is accompanied by a simulation experiment and a post-selection examination on mean income across Galicia's counties. The numerical studies confirm a good performance of our new procedure. Moreover, they reveal startling robustness to the model misspecification of a naive method to construct the confidence intervals for a mixed parameter which is in contrast to our findings for the fixed parameters.

EO292 Room Virtual R23 METHODS FOR HIGH-DIMENSIONAL AND NON-STANDARD DATA Chair: Enea Bongiorno

E0756: **Locally sparse function-on-function regression**
Presenter: **Marco Stefanucci**, University of Trieste, Italy
Co-authors: Antonio Canale, Mauro Bernardi

The focus is on functional linear regression, and specifically, we consider models for functional response and functional covariates. The literature proposes two approaches to address this situation: the concurrent functional model and the non-concurrent functional model. In the former, the value of the functional response at a given domain point depends only on the value of the functional regressors evaluated at the same domain point, while in the latter the functional covariates evaluated in each point of their domain have a non-null effect on the response in any point of its domain. To balance these two extremes, we propose a locally sparse functional regression model in which the functional regression coefficient is allowed – and not forced – to be exactly zero for a subset of its domain. We achieve this by means of a suitable basis representation of the functional regression coefficient and exploiting an overlap group-Lasso penalty for its estimation. Efficient computational strategies based on majorization-minimization

algorithms are introduced, and appealing theoretical properties in terms of models support and consistency of the proposed estimator are discussed. The empirical performance of the method is illustrated through simulation studies and an application related to human mortality.

E1215: **Evaluating the complexity of a functional time series**
Presenter: **Kwo Lik Lax Chan**, Universita degli Studi del Piemonte Orientale, Italy
Co-authors: Enea Bongiorno, Aldo Goia

Consider a functional time series taking values in a general topological space and assume that its Small-Ball Probability (SmBP) factorizes into two terms that play the role of a surrogate density and of a volume term. The latter is a means for studying the complexity of the underlying process, since it may reveal some latent feature of it. In some cases, it can be analytically specified in a parametric form: a special situation is given when the process belongs to the monomial family, like in the finite-dimensional and fractal case, for which the volumetric term has monomial form depending on the SmBP radius and a parameter named complexity index. The aim is to present some results concerning the study of a nonparametric estimator for the volume term based on a U-statistic in the beta-mixing framework. Weak consistency of this estimator is provided. In the particular case of a monomial family, it is possible to estimate the complexity index by minimizing a suitable dissimilarity measure. For this estimator asymptotic Gaussianity is shown, providing theoretical support to build confidence interval for the complexity index. A Monte Carlo simulation is carried out in order to assess the performance of the methodology for finite sample sizes. Finally, the new method is applied to detect the complexity of a real-world dataset.

E0783: **Measuring and comparing statistically the importance of terms in documents based on topic modelling**
Presenter: **Louisa Kontoghiorghes**, Kings College London, United Kingdom
Co-authors: Ana Colubi

Topic modelling is a well-known text mining technique to identify the themes covered in a set of documents. The quantification of the importance of a topic, or topic prevalence, is essential in this area. However, tracing topics in a set of documents, or time series, lacks identifiability. The proposal is to focus on keywords instead of on topics to build a new prevalence metric. The new metric is still based on topic modelling, and it involves the topics related to the considered terms. The keywords can be predetermined or automatically extracted from previous documents or topic models. The suggested approach overcomes the identifiability problem and enables us to test changes in keywords/topic prevalences statistically. Thus, as a step forward, statistical hypothesis tests in this area will be developed. Given the complexity of the involved parametric distributions, a distribution-free bootstrap approach is suggested. The methodology is applied to analyze the change of essential themes in the conference CMStatistics.

EO428 Room Virtual R24 QUANTITATIVE METHODS FOR HEALTH DISPARITIES RESEARCH Chair: Rebecca Betensky

E0819: **Disparities in missingness produce algorithmic bias: Availability of family health history in electronic medical records**
Presenter: **Melody Goodman**, NYU School of Global Public Health, United States

Prior research has shown that individuals from minority racial and ethnic groups have decreased access to and utilization of genetic services, and that these disparities cannot be attributed only to cost. Individual-level factors (e.g., awareness, knowledge, attitudes) have been identified that may contribute to disparities in access to and use of genetic services, and the importance of health care system-level factors (e.g., insurance, access to specialists, language barriers) is increasingly being recognized. To identify patients eligible for genetic services, a critical piece of information is a detailed family history which is one of the best predictors of cancer risks. While taking a family history is a key component of primary care, this information is often not adequately or routinely collected. Information about second-degree relatives and age at relatives disease diagnoses, required for risk stratification, is collected infrequently. The disparity in the availability of family history information in the EHR also contributes to bias in algorithms that aim to identify eligible patients. Our recent work suggests that Spanish-speaking and Hispanic patients are underrepresented in the pool identified by an EHR-based algorithm using structured data for identifying unaffected patients who qualify for cancer genetic services based on current guidelines from primary care clinics. We examine disparities in missingness by gender, race, ethnicity, and language.

E1301: **Simple principles to incorporate diversity, equity, and inclusion principles in biostatistics courses**
Presenter: **Scarlett Bellamy**, Drexel University, Dornsife School of Public Health, United States

As biostatistics begins to embrace conversations about improving diversity, equity, and inclusion (DEI) in the field, ideally, these perspectives will also begin to permeate outward to impact every aspect of the profession, including incorporating these principles into how we teach our trainees. By incorporating DEI principles into biostatistics pedagogy, instructors and trainees can cultivate a more holistic understanding of both historical background and current challenges in the field by enabling all students to see themselves in the content and how they might contribute to making important contributions to both statistical theory and application. Basic, practical examples will be used to introduce these concepts into courses without compromising course objectives and often without requiring additional time for these curricular enhancements, thus promoting a culture of inclusion in training environments.

E1315: **Latent variable modeling to understand characteristics associated with disparities in adult asthma**
Presenter: **Knashawn Morales**, University of Pennsylvania, United States

Asthma, a chronic treatable disease, disproportionately affects low-income and minority adults. Improving access to care and patient-provider communication is believed to be essential for better outcomes. Understanding the trajectory of improvement and subgroups that benefit from service interventions is a necessary step following the conduct of randomized trials. Latent variable modeling can capture both heterogeneous patterns in outcomes over time as well as subgroups that demonstrate more or less improvement in outcomes. We discuss how the flexible class of models applied to longitudinal data can be used to quantify these patterns while taking into account patient characteristics and social determinants of health. We then illustrate the method using pooled data from two randomized trials of service interventions for adults with uncontrolled asthma. Identifying characteristics of individuals who are less likely to respond to treatment may signal a need for more intensive interventions and be a stepping stone to reducing disparities.

EO589 Room Virtual R25 SPATIAL MODELS FOR DISEASE SURVEILLANCE Chair: Andrew Lawson

E0438: **Bayesian space-time SIR modeling of Covid19 in SC, USA**
Presenter: **Andrew Lawson**, Medical University of South Carolina, United States
Co-authors: Joanne Kim

The Covid19 pandemic has spread across the world during much of 2020. Many regions have experienced its effects. South Carolina in the USA has seen cases since early March 2020 and a primary peak in early April 2020. A lockdown was imposed on April 6th, but the lifting of restrictions started on April 24th. The daily case and death data reported by NCHS (deaths) and state health department (cases) via the New York Times GitHUB repository have been analyzed, and approaches to modeling the data are presented. Spatially-referenced Bayesian susceptible/infected/removed (SIR) models with different assumptions concerning transmission and county-neighborhood relations are examined. Prediction is also considered, and the role of asymptomatic transmission is assessed as a latent unobserved effect. Both crude daily and smoothed counts for a single time period are examined, and one step prediction is provided.

E0628: **Zero-state coupled Markov switching count models for spatio-temporal infectious disease counts**
Presenter: **Alexandra Schmidt**, McGill University, Canada

Spatio-temporal counts of infectious disease cases often contain an excess of zeros. Existing zero-inflated count models applied to such data are difficult to epidemiologically interpret in terms of how the disease spreads and do not allow for separate dynamics to affect the reemergence and persistence of the disease. As an alternative, we develop a new zero-state coupled Markov switching negative binomial model, under which the disease switches between periods of presence and absence in each area through a series of partially hidden nonhomogeneous, including random effects, Markov chains coupled between neighboring locations. When the disease is present, an autoregressive negative binomial model generates the cases with a possible 0 representing the disease being undetected. Bayesian inference and prediction are illustrated using spatio-temporal courts of dengue fever cases in Rio de Janeiro, Brazil.

E0181: **Spatial analysis of measles in Colombia using a Bayesian model that allows for risk estimation and cluster detection**
Presenter: **Ana Corberan-Vallet**, University of Valencia, Spain
Co-authors: Karen C Florez, Ingrid Carolina Marino, Jose D Bermudez

A Bayesian hierarchical Poisson model with an underlying cluster structure is applied to describe measles incidence in Colombia. Concretely, the proposed methodology provides relative risk estimates at department level and identifies clusters of disease. We also show how socio-demographic factors can be included in the model to describe disease incidence better. Since the model does not impose any spatial correlation at any level of the model hierarchy, it avoids the spatial confounding problem and provides a suitable framework to estimate the fixed-effect coefficients associated with spatially-structured covariates. This analysis will facilitate the implementation of targeted public health interventions, which are essential to restrict the expanse of the epidemic that reemerged in Venezuela in 2017

EO722 Room Virtual R26 RECENT ADVANCES IN BAYESIAN METHODS Chair: Andres Barrientos

E0356: **Autoregressive density modeling with the Gaussian process mixture transition distribution**
Presenter: **Matthew Heiner**, Brigham Young University, United States
Co-authors: Athanasios Kottas

A mixture model is developed for transition density approximation, together with soft model selection, in the presence of noisy and heterogeneous nonlinear dynamics. The model builds on the Gaussian mixture transition distribution (MTD) model for continuous state spaces, extending component means with nonlinear functions that are modeled using Gaussian process (GP) priors. The resulting model flexibly captures nonlinear and heterogeneous lag dependence when several mixture components are active, identifies low-order nonlinear dependence while inferring relevant lags when few components are active, and averages over multiple and competing single-lag models to quantify/propagate uncertainty. Sparsity-inducing priors on the mixture weights aid in selecting a subset of active lags. The hierarchical model specification follows conventions for both GP regression and MTD models, admitting a convenient Gibbs sampling scheme for posterior inference. We demonstrate the properties of the proposed model with two simulated and two real-time series, emphasizing approximation of lag-dependent transition densities and model selection. In most cases, the model decisively recovers important features. The proposed model provides a simple, yet flexible framework that preserves useful and distinguishing characteristics of the MTD model class.

E0544: **Efficient in-situ image compression through probabilistic image representation**
Presenter: **Rongjie Liu**, Florida State University, United States

Fast and effective image compression for multi-dimensional images has become increasingly important for efficient storage and transfer of massive amounts of high-resolution images. Desirable properties in compression methods include (1) high reconstruction quality at a wide range of compression rates while preserving key local details, (2) computational scalability, (3) applicability to a variety of different image types and of different dimensions and (4) ease of tuning. We present such a method for multi-dimensional image compression called Compression via Adaptive Recursive Partitioning (CARP). CARP uses an optimal permutation of the image pixels inferred from a Bayesian probabilistic model on recursive partitions of the image to reduce its effective dimensionality, achieving a parsimonious representation that preserves information. It uses a multi-layer Bayesian hierarchical model to achieve in-situ compression along with self-tuning and regularization, with just one single parameter to be specified by the user to achieve the desired compression rate. Extensive numerical experiments using a variety of datasets including 2D still images, real-life YouTube videos, and surveillance videos show that CARP compares favorably to a wide range of popular image compression approaches, including JPEG, JPEG2000, AVI, BPG, MPEG4, HEVC, AV1, and a couple of neural network-based methods.

E1008: **Mixture representations for likelihood ratio ordered distributions**
Presenter: **Michael Jauch**, Cornell University, United States
Co-authors: Andres Felipe Barrientos, Victor Pena, David Matteson

Mixture representations for likelihood ratio ordered distributions are introduced. Essentially, the ratio of two probability densities, or mass functions, is monotone if and only if one can be expressed as a mixture of one-sided truncations of the other. To illustrate the practical value of the mixture representations, we address the problem of density estimation for likelihood ratio ordered distributions. In particular, we propose a nonparametric Bayesian solution which takes advantage of the mixture representations. The prior distribution is constructed from Dirichlet process mixtures and has large support on the space of pairs of densities satisfying the monotone ratio constraint. With a simple modification to the prior distribution, we can test the equality of two distributions against the alternative of likelihood ratio ordering. We develop a Markov chain Monte Carlo algorithm for posterior inference and demonstrate the method in a biomedical application.

EO782 Room Virtual R27 SPATIAL AND SPATIO-TEMPORAL DATA SCIENCE Chair: Miryam Sarah Merk

E0510: **Model selection and model averaging for matrix exponential spatial models**
Presenter: **Osman Dogan**, University of Illinois, United States
Co-authors: Suleyman Taspinar, Ye Yang

The focus is on a model specification problem in spatial econometric models when an empiricist needs to choose from a pool of candidates for the spatial weights matrix. We propose a model selection (MS) procedure for the matrix exponential spatial specification (MESS), when the true spatial weights matrix may not be in the set of candidate spatial weights matrices. We show that the selection estimator is asymptotically optimal in the sense that asymptotically it is as efficient as the infeasible estimator that uses the best candidate spatial weights matrix. The proposed selection procedure is also consistent in the sense that when the data generating process involves spatial effects, it chooses the true spatial weights matrix with probability approaching one in large samples. We also propose a model averaging (MA) estimator that compromises across a set of candidate models. We show that it is asymptotically optimal. We further flesh out how to extend the proposed selection and averaging schemes to higher-order specifications and to the MESS with heteroskedasticity. The Monte Carlo simulation results indicate that the MS and MA estimators perform well in finite samples. We also illustrate the usefulness of the proposed MS and MA schemes in a spatially augmented economic growth model.

E0575: **An adaptive-LASSO algorithm for feature selection in functional spatiotemporal models**
Presenter: **Paolo Maranzano**, University of Bergamo, Italy
Co-authors: Alessandro Fasso, Philipp Otto

A model selection algorithm based on adaptive-LASSO regularization for spatiotemporal models is discussed. In particular, we are interested in applying a penalized likelihood feature selection procedure to functional Hidden Dynamics Geostatistical Models, or f-HDGM. These models represent the phenomenon of interest using a mixed-effects structure, in which the latent component describes the spatiotemporal dynamics and the

fixed-effects component models the interaction between the response variable and exogenous phenomena via linear regression. Model coefficients are shaped as continuous functions that vary across a functional domain. We focus on functional regression models based on B-spline basis functions for interpolation, where the number of free parameters is given by the order of the spline and the number of internal knots. We aim at identifying a robust procedure to select the subset of relevant spline basis functions used to model the relationships, employing a penalized likelihood algorithm and cross-validation to choose the best models. The proposed algorithm is applied to both simulated and real-world data. The empirical data concern the case study of hourly air pollutant concentrations observed during the lockdown period imposed in 2020 to address the spread of the COVID-19 virus in Northern Italy.

E1349: **Directional spatial autoregressive dependence in the conditional first- and second-order moments**
Presenter: **Miryam Sarah Merk**, Georg-August-Universitaet Goettingen, Lehrstuhl fuer Oekonometrie/ Statistik, Germany
Co-authors: Philipp Otto

In contrast to classical econometric approaches which are based on prespecified isotropic weighting schemes, we suggest that the spatial weighting matrix in the presence of directional dependencies should be estimated. In contrast to temporal autoregressive processes, where the direction of dependence is known in advance, we identify the direction of the spatial dependence based on different candidate neighbourhood sets. Two different types of processes, namely spatial autoregressive and spatial autoregressive conditional heteroscedastic processes, are considered and both estimated by maximum likelihood. Monte Carlo simulation results indicate that the models performance improves with sample size and with smaller neighbourhood subset sizes. Moreover, we apply the proposed approach to aerosol observations over the North Atlantic region and show that their spatial dependence matches the direction of the trade winds in this area.

EO336 Room Virtual R28 ADVANCED METHODS FOR TIME SERIES Chair: Bouchra Nasri

E0343: **Shrinkage estimators in tensor regression with change-points**
Presenter: **Mai Ghannam**, University of Windsor, Canada
Co-authors: Severien Nkurunziza

An estimation problem about the tensor coefficient in a tensor regression model with multiple and unknown number of change-points is considered. We generalize some recent findings in five ways. First, the problem studied is more general than the one in the context of a matrix parameter with multiple change-points. Second, we develop asymptotic results of the tensor estimators in the tensor regression model with unknown change-points. Third, we construct a class of shrinkage tensor estimators that encloses the unrestricted estimator (UE) and the restricted estimator (RE). Fourth, we generalize some identities, which are crucial in studying the risk dominance of tensor estimators. Fifth, we show that the proposed shrinkage estimators perform better than the UE. The additional novelty of the established results is that the dependence structure of the errors is as weak as that of L_2-mixingale tensors.

E0624: **Goodness of fit for regime-switching copula models with application to option pricing**
Presenter: **MamadouYamar Thioub**, HEC Montreal, Canada
Co-authors: Bouchra Nasri, Bruno N Remillard

Several time series are considered, and for each of them, an appropriate dynamic parametric model is fitted. This produces serially independent error terms for each time series. The dependence between these error terms is then modelled by a regime-switching copula. The EM algorithm is used for estimating the parameters and a sequential goodness of fit procedure based on Cramer-von Mises statistics is proposed to select the appropriate number of regimes. Numerical experiments are performed to assess the validity of the proposed methodology. As an example of application, we evaluate a European put-on max option on the returns of two assets. To facilitate the use of our methodology, we have built a R package HMMcopula available on CRAN.

E1127: **Spatially-coupled hidden Markov models**
Presenter: **Vianey Leos Barajas**, University of Toronto, Canada

Hidden Markov models (HMMs) provide a flexible framework to model time series data where the observation process Y is taken to be driven by an underlying latent state process Z. HMMs can accommodate multivariate processes by (I) assuming that a single state governs the M observations at time t, (ii) assuming that each observation process is governed by its own HMM, or (iii) a balance between the two, as in the coupled HMM framework. Coupled HMMs assume that a collection of M observation processes are governed by their respective M state processes, where the state process for process m at time t, depends on all other state processes at time $t-1$. We introduce spatially-coupled hidden Markov models where the state processes interact according to an imposed neighborhood structure with observations collected across N spatial locations. We outline an application to short-term forecasting of wind speed using data collected across meteorological stations.

EO274 Room Virtual R29 STATISTICAL METHODS FOR STREAMING DATA Chair: Michelle Miranda

E1572: **Quantile functional regression for high dimensional data streams**
Presenter: **Jeffrey Morris**, University of Pennsylvania, United States
Co-authors: Ye Emma Zohner

A new methodology is presented for modeling data streams in a Bayesian functional regression framework. The general strategy is to partition the data stream into serial epochs, compute the distribution of observations within each epoch that we represent in a custom quantile function space, and model as a functional object. We assess how the distributions vary over time and over subject-specific covariates. The motivating data for this methodology is intraocular pressure data collected from sensors placed on the eyes of non-human primates. This study characterizes intraocular pressure which is linked to glaucoma, a group of eye diseases that affects millions of people worldwide. Although our methodology is motivated by intraocular data streams, the methods we develop can broadly be used for functional regression on data streams to assess how subject-level and temporal factors affect the dynamic shifting distribution of the data represented in the stream.

E1619: **Online statistical inference for stochastic optimization**
Presenter: **Yichen Zhang**, Purdue University, United States
Co-authors: Xi Chen, Zehua Lai, He Li

As stochastic optimization attracts attention for a wide range of applications with complex objective functions, there is an increasing demand for uncertainty quantification of estimated parameters. We investigate the problem of statistical inference for model parameters based on gradient-free stochastic optimization methods that use only function values rather than gradients. We first present central limit theorem results for Polyak-Ruppert-averaging type gradient-free estimators. The asymptotic distribution reflects the trade-off between the rate of convergence and function query complexity. To construct valid confidence intervals based on the obtained asymptotic distribution, we further provide a general gradient-free framework for online covariance estimation and analyze the role of function query complexity in the convergence rate of the covariance estimator. This provides a one-pass computationally efficient procedure for simultaneously obtaining an estimator of model parameters and conducting statistical inference. Finally, we provide numerical experiments to verify our theoretical results and illustrate some extensions of our method for some applications.

E1718: **Damped Anderson mixing for deep reinforcement learning: Acceleration, convergence, and stabilization**
Presenter: **Linglong Kong**, University of Alberta, Canada

Anderson mixing has been heuristically applied to reinforcement learning (RL) algorithms for accelerating convergence and improving the sampling efficiency of deep RL. Despite its heuristic improvement of convergence, a rigorous mathematical justification for the benefits of Anderson mixing in RL has not yet been put forward. We provide deeper insights into a class of acceleration schemes built on Anderson mixing that improve the convergence of deep RL algorithms. The main results establish a connection between Anderson mixing and quasi-Newton methods and prove that Anderson mixing increases the convergence radius of policy iteration schemes by an extra contraction factor. The key focus of the analysis roots in the fixed-point iteration nature of RL. We further propose a stabilization strategy by introducing a stable regularization term in Anderson mixing and a differentiable, non-expansive MellowMax operator that can allow both faster convergence and more stable behavior. Extensive experiments demonstrate that our proposed method enhances the convergence, stability, and performance of RL algorithms.

EO188 Room Virtual R36 LIFETIME DATA ANALYSIS: SURVIVAL AND RELIABILITY Chair: Mariangela Zenga

E1235: **Analysis of recurrence times by the Markovian arrival process: An application to survival cancer data**
Presenter: **Pepa Ramirez Cobo**, Universidad de Cadiz, Spain
Co-authors: Rosa Lillo, Pedro de la Concepcion Morales

The Markovian arrival process (MAP) is adapted to model dependent recurrent times. New properties of the new version of the MAP concerning the correlation structures are presented. The model is applied in a medical context where the purpose is the fitting of a real bladder cancer dataset.

E1341: **Feature selection for competing risks model in high and ultra-high dimensions**
Presenter: **Marialuisa Restaino**, University of Salerno, Italy
Co-authors: Francesco Giordano

In the analysis of time-to-event data, competing risks data are encountered when individuals may fail from multiple causes and the occurrence of one failure event precludes the others from happening. Two main approaches can be used, to investigate the effects of covariates on the hazard function: cause-specific hazard (CSH) model and subdistribution hazard (SDH) model. The difference between these two approaches relies on the definition of the risk set. In CSH, subjects who experience the competing events are treated as censored, while in SDH they are included in the risk set. In many applications involving competing risks, identifying variables that have effects on CSH or SDH is a critical task. In the CSH model, screening and variable selection methods developed for Cox model can be easily extended. For the SDH approach, due to the different definitions of the risk set, naive applications of these procedures may be problematic and not suitable. This is particularly true when the number of covariates is larger than the number of observations ($n < p$ and $n << p$) and in presence of multicollinearity between covariates. Thus, the aim is to compare the performance of some existing methods for screening and selecting the most significant variables, for both CSH and SDH models, for highlighting their main advantages and disadvantages and proposing a new procedure able to identify the relevant covariates in the framework of high and ultra-high dimensions.

E1435: **Identifying the determinants of soccer coach turnover in Italian League Serie A**
Presenter: **Francesco Porro**, Universita degli Studi di Sassari, Italy
Co-authors: Marialuisa Restaino, Juan Eloy Ruiz-Castro, Mariangela Zenga

The aim is to study the causes and consequences of head coaches turnover in the top division of the Italian Soccer League (Serie A) over the period 2010-2019. Since coaches can exit from the club because of management choice or due to their own decision, we employ competing risk models, able to estimate the effects of predictors on the risk of replacement. We gather data on coach, team and club by merging information from various internet sources. Since the set of information collected includes a large number of covariates that can influence the turnover of head coaches, we identify the best set of possible predictors, by applying a non-parametric procedure based on random survival forest, and then estimating the hazard ratios associated with the possible changes by competing risks model. Furthermore, the covariates selected by the random survival forest are compared with those chosen by a traditional variable selection, i.e. the forward selection.

EO232 Room Virtual R37 COMPUTATIONAL ADVANCEMENTS IN SURVEY SAMPLING Chair: Maria Michela Dickson

E0353: **Targeted or lagged walk sampling for estimation of finite-order graph parameters**
Presenter: **Li-Chun Zhang**, University of Southampton, United Kingdom

A pure random walk in a graph is a probabilistic depth-first search algorithm that moves strictly over the edges of the graph. More generally, a walk is said to be targeted if the transition probabilities from the current node are also subject to other devices, such as random jumps or acceptance-rejection mechanisms for the proposed moves. Finally, we call a walk lagged if the transition probabilities depend on not only the current node but also a given number of previous nodes, such that it is not memoryless like either pure or targeted random walks. A novel approach is presented for estimating finite-order graph parameters based on targeted or lagged walk sampling. As an example, let y be a binary community membership indicator associated with the nodes, such as a group of pathogen carriers or a fraternal society. Let be the total number of triangles where all the three nodes belong to the community and let be that of the other triangles. The larger the ratio between and is, the higher is the transitivity among the community members compared to the overall transitivity in the graph.

E0573: **Simplified variance estimation for multistage sample surveys**
Presenter: **Guillaume Chauvet**, ENSAI-IRMAR, France

Multistage sampling designs are commonly used for household surveys. If we wish to perform longitudinal estimations, individuals from the initial sample are followed over time. If we also wish to perform cross-sectional estimations at several times, additional samples are selected at further waves and mixed with the individuals originally selected. Even in the simplest case when estimations are produced at the first time with a single sample, variance estimation is challenging since the different sources of randomness need to be accounted for, along with the needed statistical treatments (correction of unit non-response at the household and at the individual level, correction of item non-response, calibration). We consider a bootstrap solution that accounts for the features of the sampling and estimation process. This bootstrap solution is usually conservative for the true variance, in the sense that the sampling variance tends to be overestimated. The proposed bootstrap is illustrated with examples.

E0825: **Simultaneously selection of balanced and spatially balanced samples by means of simulated annealing**
Presenter: **Francesco Pantalone**, University of Southampton, United Kingdom
Co-authors: Roberto Benedetti, Maria Michela Dickson, Giuseppe Espa, Federica Piersimoni

A new sampling method for the selection of samples that are both balanced over a set of auxiliary variables and spatially balanced is proposed. A balanced sample is suited when correlation is present between the variable of interest and a set of auxiliary variables, while a spatially balanced sample is recommended when there is spatial correlation in the population. Indeed, a gain in efficiency in terms of variance of the Horvitz-Thompson estimator can be achieved in these situations by means of the aforementioned samples. The new method, which is based on a modified version of the simulated annealing, allows to face situations where a correlated set of auxiliary variables and spatial correlation are both present, since it can select samples that are balanced and spatially balanced simultaneously.

EO575 Room Virtual R38 STOCHASTIC PROCESS MODELS AND THEIR INFERENCE Chair: Michael Wiper

E0571: **A stochastic model for multiple bacteria growth curves based on the random telegraph process**
Presenter: **Michael Wiper**, Universidad Carlos III de Madrid, Spain

Co-authors: Ana Paula Palacios, J Miguel Marin

A new, stochastic model for growth curves is developed. This model is based on a time-stretched, integrated stochastic process. By design, the mean curve (assuming equilibrium) is a standard growth curve model. The underlying process is a Markov process with two states. When the associated rates are equal, the process is symmetric around the mean, in the sense that the expected skewness about the mean at any time point is zero. When the rates are different, then the process is asymmetric. As the likelihood function for this process cannot be derived, Maximum likelihood estimation cannot be used for inference purposes. Therefore, assuming a parametric mean growth curve, given a sample of growth curve data, least squares are used to estimate the curve. Then, the rates are estimated via the method of moments or approximate Bayesian computation. Furthermore, we show that we can decide whether to use the symmetric or asymmetric model using a standard test for asymmetry at a carefully chosen time point. The model is illustrated using both simulated data and a real data set of listeria growth curves.

E1421: **Controlled branching processes: Estimation based on ABC-SMC methodology**

Presenter: **Miguel Gonzalez Velasco**, University of Extremadura, Spain

Co-authors: Carmen Minuesa Abril, Ines M del Puerto

The focus is on the estimation of the posterior distribution of the main parameters of a controlled branching process (CBP) without explicit likelihood calculations. Specifically, we focus on the case where we have no prior knowledge of the maximum number of offspring that an individual can produce. Our approach has two steps. In the first stage, we estimate the posterior distribution of the maximum progeny per individual using an approximate Bayesian computation (ABC) algorithm for model choice with the raw data and based on sequential importance sampling. In the second step, using the values simulated in the previous stage, we estimate the posterior distribution of the main parameters of a CBP by applying the rejection ABC algorithm with an appropriate summary statistic and a post-processing adjustment. We show the accuracy of the proposed methodology via simulated examples and via real data from models that incorporate a carrying capacity, in both cases making use of the statistical software R.

E1482: **A gamma process application to microorganisms population modelling**

Presenter: **J Miguel Marin**, University Carlos III, Spain

Co-authors: Michael Wiper, Fabrizio Ruggeri

Models are proposed for longitudinal data with a multilevel structure in the context of multiple experiments with microorganisms. We use a gamma process with an embedded structure of splines. We analyze a bunch of experimental data and obtain predictions of growing under different conditions.

EO342 Room Virtual R39 RECENT DEVELOPMENTS IN STATISTICAL NETWORK ANALYSIS Chair: Jonathan Stewart

E0539: **The importance of being correlated: Implications of dependence in joint spectral inference across multiple networks**

Presenter: **Vince Lyzinski**, University of Maryland, College Park, United States

Co-authors: Konstantinos Pantazis, Avanti Athreya, Jesus Arroyo, William Frost, Evan Hill

Spectral inference on multiple networks is a rapidly-developing subfield of graph statistics. Recent work has demonstrated that joint, or simultaneous, spectral embedding of multiple independent networks can deliver more accurate estimation than individual spectral decompositions of those same networks. Such inference procedures typically rely heavily on independence assumptions across the multiple network realizations, and even in this case, little attention has been paid to the induced network correlation in such joint embeddings. We present a generalized omnibus embedding methodology and provide a detailed analysis of this embedding across both independent and correlated networks, the latter of which significantly extends the reach of such procedures. We describe how this omnibus embedding can itself induce correlation, leading us to distinguish between inherent correlation – the correlation that arises naturally in multisample network data – and induced correlation, which is an artifice of the joint embedding methodology. We examine how induced and inherent correlation can impact inference for network time-series data, and we provide network analogues of classical questions such as the effective sample size for more generally correlated data. Further, we show how an appropriately calibrated generalized omnibus embedding can detect changes in real biological networks that previous embedding procedures could not discern.

E0893: **On the asymptotics of temporal motif estimation via sampling**

Presenter: **Eric Kolaczyk**, Boston University, United States

Co-authors: Xiaojing Zhu

Similarly to motifs (small subgraph patterns) in static networks, temporal motifs are the fundamental building blocks for temporal structures in dynamic networks consisting of a set of vertices and a collection of timestamped interaction events, i.e., temporal edges, between vertices. Temporal motifs are defined as classes of isomorphic induced subgraphs on sequences of temporal edges, considering both edge ordering and duration. Several methods have been designed to count the occurrences of temporal motifs in dynamic networks, with recent work focusing on estimating the count under various sampling schemes along with concentration properties. However, little attention has been given to the asymptotics that result. We provide conditions for the consistency and the asymptotic normality of the Horvitz-Thompson type of estimator in an edge sampling framework, which can be used to construct confidence intervals and hypothesis testing for the temporal motif count in the sampling model. We also discuss these conditions under various stochastic models for dynamic networks with temporal edges from the class of multivariate counting processes.

E1292: **Estimating random graph models from observed subgraphs**

Presenter: **Jonathan Stewart**, Florida State University, United States

The statistical analysis of sampled network data has been an important and underdeveloped topic in the field of statistical network analysis. We consider the problem of estimating a random graph model from only an observed subgraph. Current approaches make either restrictive assumptions about the dependence structure of the random graph model or require missing data methods that are not scalable to larger graphs. Recent developments are presented that design scalable estimation methodology for estimating parameter vectors of increasing dimension for population models of random graphs with dependent edges based on an observed subgraph. Our methodology designs observation processes that exploit the dependence structure of models in order to ensure sufficient information is contained within the sampled subgraph to facilitate scalable estimation. We show that common observation processes for sampling networks produce observed subgraphs which contain sufficient information for estimating models with triadic dependence. We conclude by elaborating sufficient conditions under which we can obtain non-asymptotic bounds on the statistical error of our estimators.

EO513 Room Virtual R40 FALSE CONFIDENCE, UNVERIFIABLE ASSUMPTIONS: FOUNDATIONS MATTER Chair: Peter Grunwald

E0748: **How to think about model assumptions**

Presenter: **Christian Hennig**, University of Bologna, Italy

The starting point is the apparently popular idea that in order to do (frequentist) model-based inference we need to believe that the model is true, and the model assumptions need to be fulfilled. We will argue that this is a misconception. Models are, by their very nature, not "true" in reality. Mathematical results secure favourable characteristics of inference in an artificial model world in which the model assumptions are fulfilled. For using a model in reality we need to ask what happens if the model is violated in a "realistic" way. One key approach is to model a situation in which certain model assumptions of the method that we want to apply, are violated, in order to find out what happens then. This, somewhat inconveniently, depends strongly on what we assume, how the model assumptions are violated, whether we make an effort to check them, how we do that, and

what alternative actions we take if we find them wanting. We will discuss what we know and what we can't know regarding the appropriateness of the models that we "assume", and how to interpret them appropriately, including new results on conditions for model assumption checking to work well, and on untestable assumptions.

E0863: **False confidence, imprecise probability, and valid statistical inference**
Presenter: **Ryan Martin**, North Carolina State University, United States

Statistical inference aims to quantify uncertainty about unknowns based on data. To formalize this, an inferential model (IM) is a function that maps data, etc., to a capacity on the parameter space assigning data-dependent degrees of belief to assertions about the unknowns; this covers Bayes, fiducial, and other distributional inference approaches. Important questions include: what statistical properties should an IM satisfy, and what do these statistical properties imply about the mathematical structure of its capacity? We will define a "validity" property which, among other things, implies strong frequentist error rate control. Then we will summarize two recent results saying that (a) an IM whose capacity is a precise/additive probability suffers from false confidence and, therefore, cannot be valid, and (b) validity can be achieved by IMs whose capacities belong to a simple class of imprecise/non-additive probabilities, namely, possibility measures. We will end with illustrations and a discussion of practical implications and open questions.

E1156: **Evidence and the optional continuation principle**
Presenter: **Peter Grunwald**, CWI and Leiden University, Netherlands

How much evidence do the data give us about one hypothesis versus another? The standard way to measure evidence is still the p-value, despite a myriad of problems surrounding it. One central such problem is its inability to deal with optional continuation (a weaker and, we argue, more urgent requirement than optional stopping) and its dependence on unknowable counterfactuals. The E-value is a notion of evidence which overcomes these issues. When both hypotheses are simple, the E-value is a likelihood ratio (LR) - nowadays the standard notion of probabilistic evidence in courts of law. When there is a null hypothesis and it is simple, the E-value coincides with the Bayes factor, the notion of evidence preferred by Bayesians. But while nonparametric hypotheses and/or lack of crisp alternatives pose difficulties for LRs and Bayes factors, one can still design useful E-values for them.

EC870 Room Virtual R35 CONTRIBUTIONS IN COMPOSITIONAL DATA ANALYSIS Chair: Karel Hron

E0287: **Compositional models for mutational signature analysis**
Presenter: **Lena Morrill**, University of Cambridge, United Kingdom
Co-authors: Florian Markowetz

Mutational processes leave their imprint on the DNA in the form of inherited mutations from ancestor cells, accumulating over time. We are interested in a particular type of mutation called a copy number change, in which big sections of the genome are gained, lost, or re-arranged. Our group previously described a method to extract copy number mutational signatures. The crucial quantities of interest are the "signature exposures", which indicate the fraction of copy number changes attributable to each mutational process. These quantities are compositional, in that they sum up to one sample-wise, and we can only analyse their relative abundances. We present a partial ILR mixed-effects model for non-zero exposures and a correlated Bernoulli model for zero exposures to determine if the mutational signature spectrum changes in a coordinated way between two conditions. We apply these models to address fundamental, but unanswered, questions in high grade serous ovarian carcinoma (HGSOC), one of the most deadly cancer types. We answer the question of whether early-stage HGSOC differs from late-stage HGSOC (it does), whether primary HGSOC differs from relapsed HGSOC (it does not), and how HGSOC samples with whole-genome duplication, which appears to be crucial in cancer progression, differ from diploid genomes.

E1471: **Approximation of density functions using compositional splines with optimal knots**
Presenter: **Jitka Machalova**, Palacky University, Czech Republic
Co-authors: Karel Hron

Probability density functions result in practice frequently from the aggregation of massive data, and their further statistical processing is thus of increasing importance. However, the specific properties of density functions prevent from analyzing a sample of densities directly using tools of functional data analysis. Moreover, it is not only about the unit integral constraint, which results from the representation of densities within the equivalence class of proportional positive-valued functions, but also about their relative scale, which emphasizes the effect of small relative contributions of Borel subsets to the overall measure of the support. For practical data processing, it is popular to approximate first the input (discrete) data with a proper spline representation. In this case, the compositional splines, a new class of B-splines in the Bayes space, are suitable for the representation of density functions. The aim is to show the use of the compositional splines, especially the optimal choice of number and position of spline knots is discussed. Accordingly, the original densities are expressed as real functions using the centred log-ratio transformation, and optimal smoothing splines with a new B-spline basis honoring the resulting zero-integral constraint are developed.

E0866: **Correlations at the margins cannot be rescued by estimation**
Presenter: **Gregory Gloor**, University of Western Ontario, Canada

Count compositional data result from high throughput sequencing datasets generated by platforms that have an upper bound on the number of reads delivered such as the Illumina instruments. Correlation in compositional data is properly determined using ratio information, but recently it was used modelling to show that compositional association near the low count margin was unreliable. In this modelling the discrete nature of the data caused the ratio information to be undetermined, leading to the loss of a known association. We attempt to rescue these associations using several methods, including naive priors, imputation, amalgamation, and probabilistic modelling. Results show that low count associations cannot be recovered by any of these methods. These results further show that low count features in compositional data are unreliable and that findings that are dependent wholly or partly on low count features are suspect.

EC873 Room K2.41 (Hybrid 09) GRAPHICAL MODELS AND NETWORKS Chair: Ana Belen Ramos-Guajardo

E1282: **Estimating the normalizing constant in Bayesian networks**
Presenter: **Fritz Bayer**, ETH Zurich, Switzerland
Co-authors: Giusi Moffa, Niko Beerenwinkel, Jack Kuipers

Bayesian Networks are probabilistic graphical models that can efficiently represent dependencies among random variables. Missing data and hidden variables require calculating the probability of a subset of the random variables, the so-called normalizing constant. While knowledge of the normalizing constant is crucial for various problems in statistics and machine learning, its exact computation is usually not possible for categorical variables due to the NP-hardness of this task. We develop a divide-and-conquer approach using the graphical properties of Bayesian networks to split the computation of the normalizing constant into sub-calculations of lower complexity. Exploiting this property, we present an efficient and scalable algorithm for estimating the normalizing constant for categorical variables. Our novel method displays superior performance in a benchmarking study, where we compare it against state-of-the-art approximate inference methods. As an immediate application, we demonstrate how we can use the normalizing constant to classify incomplete data against Bayesian network clusters and use this approach for identifying the cancer subtype of kidney cancer sample genotypes. The proposed scheme enables the efficient application of Bayesian networks on incomplete data or hidden variables and is presented as a general framework that can be generalized to other exact and approximate inference schemes.

E1553: **Bayesian mixed-effect models for independent dynamic social network data**
Presenter: **Fabio Vieira**, Tilburg University, Netherlands
Co-authors: Joris Mulder, Roger Leenders, Daniel McFarland

The development of technological devices and communication applications has changed the way humans interact and provide vast amounts of data. As a result, relational event data or timestamped social network data, have been increasingly available over the years. Late developments in statistical modeling of such data focus on methods based on log-linear models. The goal is to model the rates of interactions among actors in a social network via actor covariates and network statistics. The use of survival analysis concepts has allowed the treatment of temporal evolution in social networks. Therefore, more flexible models may be developed with the goal of unveiling the effects driving the network dynamics. We propose a new Bayesian hierarchical modeling approach of independent relational event sequences. This model allows inferences at the actor level, which are useful in understanding which effects guide actors preferences in social interactions. We also present Bayes factor methods for hypothesis testing in this class of models. In addition, a new empirical Bayes factor to test random-effect structures is developed. In this test, we let the prior be determined by the data, alleviating the issue of employing improper priors in Bayes factors and thus preventing the use of ad-hoc choices in absence of prior information, which makes this test quite generally applicable. We use data of classroom interactions among high school students to illustrate the proposed methods.

E1434: **Network regression and supervised centrality estimation**
Presenter: **Junhui Jeffrey Cai**, University of Pennsylvania, United States
Co-authors: Dan Yang, Wu Zhu, Haipeng Shen, Linda Zhao

Directed networks play a crucial role in our lives and have implications for individuals' behavior. The node's position in the network, usually captured by the centrality, is an important intermediary of network effects, and is often incorporated in a regression model to elucidate the network effect on the outcome variable of interest. In empirical studies, researchers often adopt a two-stage procedure – first estimate the centrality from the observed network and then employ the estimated centrality in regression. Despite the prevalent adoption of such a two-stage procedure, it fails to incorporate the errors from the observed network and lacks valid inference. We first propose a unified inferential framework that combines the network error model and the regression on centrality model, under which we prove the shortcoming of the two-stage in estimating the centrality and demonstrate the consequent undesirable effect in the outcome regression. We then propose a novel supervised network centrality estimation (SuperCENT) methodology that simultaneously combines the information from the two essential models. The proposed method always provides superior estimates of the centrality and the true underlying network over the two-stage procedure, and produces better network effect estimation when the observational error of the network is severe. We further derive the distribution of the centrality and network effect, which can be used to construct valid confidence intervals.

EG061 Room K0.18 (Hybrid 03) CONTRIBUTIONS IN CLUSTERING COMPLEX DATA Chair: Faicel Chamroukhi

E1651: **Multilevel latent class models for cross-classified data**
Presenter: **Silvia Columbu**, University of Cagliari, Italy
Co-authors: Jeroen Vermunt

Latent class and mixture models have been extended to deal with multilevel datasets, for example, when children are nested within schools. One very useful version is an approach in which one has latent classes (or mixture components) at different levels (thus for both schools and children). These models can be estimated by maximum likelihood using a special implementation of the EM algorithm. However, sometimes the nesting consists of multiple higher levels which are not hierarchically linked, but instead cross-classified, for example, when children are nested within both schools and neighborhoods. We show how such a situation can be dealt with by having a separate set of mixture components for each of the crossed classifications. Unfortunately, given the intractability of the derived loglikelihood, the EM algorithm can no longer be used in the estimation process. We, therefore, propose an approximate estimation of this model using a stochastic version of the EM algorithm similar to Gibbs sampling.

E1524: **Mixtures-of-experts with functional predictors**
Presenter: **Faicel Chamroukhi**, Caen University, Lab of Mathematics LMNO, France
Co-authors: Nhat-Thien Pham, Van Ha Hoang, Geoffrey McLachlan

Mixtures-of-experts (ME) modeling is a popular and successful framework in prediction and clustering of heterogeneous observations with associated vectorial covariates. We consider the model-based clustering and prediction with ME models in the presence of functional covariates, and present extensions to the functional data context. The new functional ME (FME) model allows to accurately capturing complex nonlinear relationships between a scalar response and a set of predictors $\{X(t), t \in \mathbb{T} \subset \mathbb{R}\}$, which are observed continuously (i.e, over time for time series), from entire functions and are potentially noisy, and the pair $(\{X(t)\}, Y)$ is governed by an unknown hidden structure Z. We provide sparse and interpretable functional representations of the FME model, thanks to Lasso-like regularizations, notably on the derivatives of the underlying functional parameters of the model, projected onto a set of continuous basis functions. We develop dedicated EM algorithms for the regularized maximum-likelihood parameter estimation. The good performance of the proposed FME model and the developed algorithms is shown in simulated scenarios and via application to some real data sets.

E1577: **Model-based clustering of dual-energy CT images for tumor analysis**
Presenter: **Segolene Brivet**, McGill University, Canada
Co-authors: Faicel Chamroukhi, Mark Coates, Reza Forghani, Peter Savadjiev

Computed Tomography (CT) scans are commonly used for the evaluation of head and neck cancer but automatic tumor analysis can be challenging on conventional CT. We use an advanced form of CT known as Dual-Energy CT (DECT) or spectral CT. DECT may be viewed as a 4D image of a patient: a 3D body volume over a range of spectral attenuation levels. The latter dimension provides, for each voxel, a decay curve representing energy-dependent changes in attenuation that enables tissue characterization beyond what is possible with conventional CT. We propose a clustering method that uses spectral tissue characteristics to segment the image into areas with consistent contours and high-quality features. The clusters could be used in tumor segmentation or cancer outcome prediction. We construct functional mixture models that specifically integrate spatial context in mixture weights, with mixture component densities being constructed upon the energy decay curves as functional observations. This accommodates the spectral energy curve nature of the data. We design unsupervised clustering algorithms by developing dedicated expectation-maximization (EM) algorithms to estimate the maximum likelihood of the model parameters. The method was evaluated on 90 head and neck DECT scans, each containing a tumor contoured by radiologists. Our algorithm performs well in clustering the anatomical tumor region, as demonstrated by comparing its coverage with the ground truth contour.

CI014 Room K E. Safra (Multi-use 01) ADVANCES IN MACRO AND FINANCE (VIRTUAL) Chair: Martin Wagner

C0728: **Robustifying inference of DSGE models estimated by filtering methods**
Presenter: **Martin M Andreasen**, Aarhus University, Denmark

Some challenges are discussed, which are related to estimating potentially nonlinear DSGE models by filtering methods. We also discuss simple ways to detect model misspecification. To make the existing (quasi) likelihood-based estimators more robust, we augment the (quasi) likelihood function by a set of unconditional GMM moment conditions. For a simulation study, we consider a standard New Keynesian model tailored to

match the yield curve in the US. The simulation study shows that this penalized (quasi) likelihood function approach delivers more robust estimates when the model is misspecified than the standard (quasi) likelihood approach.

C0775: **Large order-invariant bayesian vars with stochastic volatility**
Presenter: **Joshua Chan**, Purdue University, United States

Many popular specifications for Vector Autoregressions (VARs) with multivariate stochastic volatility are not invariant to the way the variables are ordered due to the use of a Cholesky decomposition for the error covariance matrix. We show that the order invariance problem in existing approaches is likely to become more serious in large VARs. We propose the use of a specification that avoids the use of this Cholesky decomposition. We show that the presence of multivariate stochastic volatility allows for the identification of the proposed model and prove that it is invariant to ordering. We develop a Markov Chain Monte Carlo algorithm which allows for Bayesian estimation and prediction. In exercises involving artificial and real macroeconomic data, we demonstrate that the choice of variable ordering can have non-negligible effects on empirical results. In a macroeconomic forecasting exercise involving VARs with 20 variables, we find that our order-invariant approach leads to the best forecasts and that some choices of variable ordering can lead to poor forecasts using a conventional, non-order invariant, approach.

CO050 Room Virtual R18 HETEROGENEOUS AND NONLINEAR DYNAMICS IN PANELS Chair: Peter Pedroni

C1726: **The macro stabilization role of migrants FDI contributions and wage remittances**
Presenter: **Borel Ntsafack**, Center for Development Economics, Cameroon
Co-authors: Patrick NCho, Peter Pedroni

Migration has long been understood to be responsive to international wage and household welfare differentials and therein to serve as an important mechanism to raise income levels for developing economies. What has been less understood is the potential short-run macro stabilization role that remittances, as well as FDI contributions from migrants, can play, both for the countries of origin as well as the host countries. To investigate this, we employ a heterogeneous panel structural VAR approach to estimating the stabilization effect on aggregate income fluctuations due to remittances and FDI contributions associated with migrants. In particular, we find that the macro stabilization role is strongest for economies that are characterized as informal, poorly resource endowed and characterized as fragile states. We also demonstrate how the technique can be used to forecast the quantitative implications of remittances and migrant FDI contributions for individual countries that have large migrant outflows, but for which high-quality data is not available. Examples include Afghanistan, Myanmar, South Sudan, Venezuela and Yemen. In addition, we use the technique to explore the potential enhanced stabilization benefits from migrant contributions for African countries that join currency unions.

C1724: **Monetary policy effectiveness the case of an economic and monetary Union with a fixed exchange rate**
Presenter: **Patrick NCho**, Center for Development Economics, Cote d'Ivoire

Low-income countries are generally associated with poor financial market development which leads to inefficiency of the monetary policy implemented by central banks. To investigate the extent to which this can affect the real economic activity of low-income countries through the lending rate of commercial banks, we use a heterogeneous panel structural VAR approach to estimate the dynamic responses of GDP, net exports and inflation to a monetary policy shock. We also investigate the impulse responses of lending rates to different structural shocks. Using the same technique, we find in particular that interest rates in low-income countries are rigid in the downward direction, even conditional on a perfect transmission of monetary policy. We also employ the technique to access the role of the exchange rate regime in explaining the difference between low-income countries with fixed exchange rates (for example in West African Economic and Monetary Union) compared to those with flexible exchange rates. Furthermore, we conclude that for the WAEMU countries, loosening monetary policy by increasing the Central bank refinancing will mainly lead to an increase in public borrowing to the detriment of the private sector.

C1720: **Using heterogeneous panels to estimate nonlinear VAR dynamics**
Presenter: **Peter Pedroni**, Williams College, United States

A new technique is developed for estimating nonlinear VAR representations of transition dynamics in heterogeneous time series panels. Specifically the technique uses a two-step approach by first estimating a heterogenous sample distribution of linear approximations for the dynamics and then using the cross-sectional variation to estimate state-dependent nonlinear expansions. The approach is thereby able to exploit the greater variation in historical experiences that are typically present in multi-country panels relative to single country time series in order to estimate a state dependency function. Monte Carlo simulations show promising small sample performance. The technique is illustrated with the estimation dynamic fiscal multipliers which are represented as a vector function of the GDP growth rate and fiscal expenditure as a share of GDP these vary over the phases of the business cycle.

CO442 Room Virtual R32 DEVELOPMENTS IN CRYPTOCURRENCY AND BLOCKCHAIN Chair: Marco Lorusso

C0822: **Conditional tail dependence between major cryptocurrencies and other major assets**
Presenter: **Pierangelo De Pace**, Pomona College, United States
Co-authors: Jayant Rao

A daily dataset of four major cryptocurrency prices, four stock indices, and three commodity prices between the beginning of 2015 and the end of 2020 is used. We empirically examine the time-varying properties of the comovement between cryptocurrency price returns and the other asset price returns. We then analyze the conditional tail dependence of their price returns by adopting a time-varying conditional copula modelling approach. Our results are heterogeneous. We show that the residual correlations between cryptocurrency and other assets are generally low and close to zero. We find the tail dependence to be small along with the sample and to experience a visible increase during the first wave of the Covid-19 pandemic.

C0995: **Trading and arbitrage with stablecoins**
Presenter: **Gerald Dwyer**, Clemson University, United States
Co-authors: Peiyun Jin

Trades data and order book from 13 exchanges are examined to study the arbitrage opportunities in stablecoins. Using snapshots of the order book data, we find that occasional prices far from the arbitrage-free prices (flash crashes) are real. Flash crashes invariably are associated with unusually large aggregate trading volume. We also examine arbitrage profitability and fee structure for stablecoin-USD and stablecoin-stablecoin. The results show that after subtracting fees, there are still positive arbitrage profits. We also compare the efficiency of Decentralized Exchanges (DEXs) and Centralized Exchanges (CEXs). Our results are consistent with DEXs and CEXs being equally efficient.

C1095: **A factor model for cryptocurrency returns**
Presenter: **Daniele Bianchi**, Queen Mary University of London, United Kingdom
Co-authors: Mykola Babiak

The factor structure of a large cross-section of daily returns on digital assets is investigated through the lens of an Instrumented Principal Component Analysis (IPCA). We show that a model with three latent factors and time-varying factor loadings significantly outperforms a benchmark six-factor model with traded, observable, factors: the in-sample (out-of-sample) R^2 from the IPCA stands at 17% (2.8%) for daily returns, against a benchmark 8% (0.2%) obtained from a model with observable risk factors taken from the existing literature. By looking at the characteristics that significantly matter for the dynamics of the latent factors, we provide robust evidence that risk premiums for digital assets are primarily driven by liquidity,

volatility, downside risk, and the market beta. The results hold both from a sample of daily returns from December 2nd 2016 to July 9th 2021 and across different sub-samples. Due to the inherent differences in cryptocurrencies and the novel and emergent status of digital assets as a form of investment, this research will be relevant to a broad audience; from market participants seeking different sources of returns and diversification, to regulators wishing to understand the role of digital assets, and to academics searching for new insights into the drivers and risks in cryptocurrency markets.

CO677 Room Virtual R33 COPULA-BASED MULTIVARIATE TIME SERIES MODELS Chair: Aleksey Min

C0692: **Change-point problems for multivariate time series using pseudo-observations**
Presenter: **Bruno N Remillard**, HEC Montreal, Canada
Co-authors: Bouchra Nasri, Tarik Bahraoui

It is shown that under weak assumptions, the change-point tests designed for independent random vectors can also be used with pseudo-observations for testing change-point in the joint distribution of non-observable random vectors, the associated copula, or the margins, without modifying the limiting distributions. In particular, change-point tests can be applied to the residuals of stochastic volatility models or conditional distribution functions applied to the observations, which are prime examples of pseudo-observations. Since the limiting distribution of test statistics depends on the unknown joint distribution function or its associated unknown copula when the dimension is greater than one, we also show that iid multipliers and traditional bootstrap can be used with pseudo-observations to approximate p-values for the test statistics. Numerical experiments are performed in order to compare the different statistics and bootstrapping methods. Examples of applications to change-point problems are given.

C1089: **Conditional empirical copula processes and generalized measures of association**
Presenter: **Alexis Derumigny**, Delft University of Technology, Netherlands
Co-authors: Jean-David Fermanian

The purpose is to study the weak convergence of conditional empirical copula processes, when the conditioning event has a nonzero probability. The validity of several bootstrap schemes is stated, including the exchangeable bootstrap. We define general - possibly conditional - multivariate measures of association and their estimators. By applying our theoretical results, we prove the asymptotic normality of some estimators of such measures. We illustrate our results with financial data.

C0495: **Detecting departures from meta-ellipticity for multivariate stationary time series**
Presenter: **Aleksey Min**, Technical University of Munich, Germany
Co-authors: Axel Buecher, Miriam Jaser

A test for detecting departures from meta-ellipticity for multivariate stationary time series is proposed. The large sample behavior of the test statistic is shown to depend in a complicated way on the underlying copula as well as on the serial dependence. Valid asymptotic critical values are obtained by a bootstrap device based on subsampling. The finite-sample performance of the test is investigated in a large-scale simulation study, and the theoretical results are illustrated by a case study involving financial log returns.

CO663 Room Virtual R34 HIGH-DIMENSIONALITY AND SPARSITY Chair: Anders Kock

C1313: **High-dimensional generalized least squares**
Presenter: **Kaveh Salehzadeh Nobari**, Lancaster University, United Kingdom
Co-authors: Anders Kock, Alex Gibberd

The finite sample properties of the GLS-LASSO estimator for high-dimensional regressions with potentially autocorrelated errors are studied. We further assess the performance of a feasible GLS-LASSO estimator and establish non-asymptotic oracle inequalities for estimation accuracy within this framework. We consider settings where the number of parameters is significantly greater than the sample size. We then illustrate the usefulness of the proposed estimators for application in high-dimensional temporal disaggregation, where disaggregation methods are used to disaggregate a time-series using an $n \times p$ indicator series matrix with $p \gg n$. Finally, we conduct a Monte Carlo Study to assess the finite sample performance of the feasible GLS-LASSO estimators in terms of estimation and variable selection accuracy using an array of error structures.

C1692: **Generalized linear models with structured sparsity estimators**
Presenter: **Mehmet Caner**, North Carolina State University, United States

Structured sparsity estimators in Generalized Linear Models are introduced. Structured sparsity estimators in the least-squares loss have been previously introduced. The proofs exclusively depended on their fixed design and normal errors. We extend the results to debiased structured sparsity estimators with Generalized Linear Model-based loss with random design and non-sub Gaussian data. Structured sparsity estimation means penalized loss functions with a possible sparsity structure in a norm. These norms include weighted group lasso, lasso and norms generated from convex cones. The contributions are fivefold: 1. We generalize the existing oracle inequality results in penalized Generalized Linear Models by proving the underlying conditions rather than assuming them. One of the key issues is the proof of a sample one-point margin condition and its use in an oracle inequality. 2. The results cover even non-sub-Gaussian errors and random regressors. 3. We provide a feasible weighted nodewise regression proof that generalizes the results in the literature from a simple l1 norm usage to norms generated from convex cones. 4. We realize that norms used in feasible nodewise regression proofs should be weaker or equal to the norms in penalized Generalized Linear Model loss. 5. We can debias the first step estimator via getting an approximate inverse of the singular-sample second-order partial derivative of Generalized Linear Model loss.

C1545: **Bridging factor and sparse models**
Presenter: **Ricardo Masini**, Princeton University, United States
Co-authors: Marcelo Medeiros, Ricardo P Masini

Factor and sparse models are two widely used methods to impose a low-dimensional structure in high-dimension. They are seemingly mutually exclusive. We propose a simple lifting method that combines the merits of these two models in a supervised learning methodology that allows us to efficiently explore all the information in high-dimensional datasets. The method is based on a flexible model for panel data, called factor-augmented regression model with both observable or latent common factors, as well as idiosyncratic components as high-dimensional covariate variables. This model not only includes both principal component (factor) regression and sparse regression as specific models but also significantly weakens the cross-sectional dependence and hence facilitates model selection and interpretability. The methodology consists of three steps. At each step, the remaining cross-section dependence can be inferred by a novel test for covariance structure in high-dimensions. We developed asymptotic theory for the factor-augmented sparse regression model and demonstrated the validity of the multiplier bootstrap for testing high-dimensional covariance structures. This is further extended to testing high-dimensional partial covariance structures.

CG033 Room Virtual R30 CONTRIBUTIONS IN FINANCIAL ECONOMETRICS I Chair: Christos Savva

C1314: **Stock returns predictability with unstable predictors**
Presenter: **Simon Price**, University of Essex, United Kingdom
Co-authors: George Kapetanios, Fabio Calonaci

The predictability of US stock returns is re-examined. Theoretically, well-founded models predict that stationary combinations of I(1) variables

such as the dividend or earnings to price ratios or the consumption/asset/income relationship often known as CAY may predict returns. However, there is evidence that these relationships are unstable, and that allowing for discrete shifts in the unconditional mean (location shifts) can lead to greater predictability. It is unclear why there should be a small number of discrete shifts and we allow for more general instability in the predictors, characterised by smooth variation. This method, which removes non-stationarity and reduces persistence in near unit root processes may be seen as an alternative to the popular IVX methods where there is strong persistence in the predictor. We apply this methodology to the three predictors mentioned above between 1952 and 2019, including the financial crisis but excluding the Covid pandemic, and find that modelling smooth instability improves predictability and forecasting performance and tends to outperform discrete location shifts, whether identified by in-sample Bai-Perron tests or Markov-switching models.

C0302: **Risk-return tradeoff in international stock returns: Skewness and business cycles**

Presenter: **Christos Savva**, Cyprus University of Technology, Cyprus

Co-authors: Henri Nyberg

The fundamental risk-return relation with a flexible regime-switching model is examined by combining the impact of skewness and business cycle regimes in stock returns. The key methodological and empirical findings point out the need for our highly nonlinear and non-Gaussian model to get a reliable picture of the risk-return relationship. With an international dataset of major countries to global financial markets, we find that accounting especially for skewness patterns leads to the expected positive risk-return relation, which is importantly also maintained over different business cycle conditions.

C1767: **Rates of return on stock prices: Impact of ESG measures**

Presenter: **Piotr Jaworski**, University of Warsaw, Poland

The aim is to examine the impact of the environmental social and governance (ESG) measures on the rates of return on stock prices of non-financial institutions according to the size of the company and sector divisions. The hypotheses are as follows: ESG risk has got a strong negative impact on the rates of return on the stock prices of the non-financial institutions. The reaction of the rates of return is varied in different sectors and stronger for bigger companies. Panel event models were used to verify these hypotheses. The study used data from the Thomson Reuters Database for the period 2010-2021. The analysis was based on papers and reports on COVID-19, ESG factors and their impact on the rates of return on stock prices changes. The analysis has been made for particular sectors according to the Thomson Reuters division and the size of companies measured by the assets value and the capitalization.

CG009 Room Virtual R31 CONTRIBUTIONS IN MONETARY POLICIES Chair: Nektarios Michail

C1477: **Forward guidance and conventional monetary policy shocks: Identification and assessment**

Presenter: **Mirela Sorina Miescu**, Lancaster University, United Kingdom

In January 2012 the Federal Reserve has introduced forward guidance in the form of a published policy-rate path. We exploit this change in the conduct of forward guidance and the zero lower bound (ZLB) constraints to identify a forward guidance shock and a conventional monetary policy shock. The two shocks produce conventionally signed impulse responses of output, prices, and financial variables. In line with the recent theories that help resolve the forward guidance puzzle, we find attenuated effects of forward guidance compared to the standard monetary policy shock.

C0212: **On the effectiveness of quantitative easing in the US**

Presenter: **Nektarios Michail**, Cyprus University of Technology, Cyprus

The transmission channels through which asset purchases are supposed to affect the economy are examined using US data in a Bayesian VAR setup. After distinguishing between GDP components, to account for the increase in government spending during the quantitative easing (QE) period, the results offer only weak support for the existence of a portfolio rebalancing channel. No support is found for the uncertainty or expectations channels. Overall, asset purchases appear to have had some impact on funding conditions, but little (if any) impact on the real economy. Hence, QE is more relevant in its interplay with fiscal policies, notably because it lowers the cost of debt.

C0850: **COVID-19 effects on the Canadian term structure of interest rates**

Presenter: **Marzia Cremona**, Universita Laval, Canada

Co-authors: Federico Severino, Eric Dadie

In Canada, COVID-19 pandemic triggered exceptional monetary policy interventions by the central bank, which in March 2020 made multiple unscheduled cuts to its target rate. The aim is to assess the extent to which Bank of Canada interventions affected the determinants of the yield curve. By applying Functional Principal Component Analysis to the term structure of interest rates we find that, during the pandemic, the long-run dependence of level and slope components of the yield curve is unchanged with respect to previous months, although the shape of the mean yield curve completely changed after target rate cuts. Bank of Canada was effective in lowering the whole yield curve and correcting the inverted hump of previous months, but it was not able to reduce the exposure to already existing long-run risks.

Sunday 19.12.2021 17:35 - 19:15 Parallel Session K – CFE-CMStatistics

EI018 Room K E. Safra (Multi-use 01) CAUSAL INFERENCE WITH MACHINE LEARNING (VIRTUAL) Chair: Xavier de Luna

E0250: **Improved doubly robust inference for treatment effect heterogeneity using nonparametric and high-dimensional models**
Presenter: **Joseph Antonelli**, University of Florida, United States
Co-authors: Heejun Shin

A doubly robust approach is proposed to characterizing treatment effect heterogeneity in observational studies. We utilize posterior distributions for both the propensity score and outcome regression models to provide valid inference on the conditional average treatment effect even when high dimensional or nonparametric models are used. We show that our approach provides conservative inference in finite samples or under model misspecification, and provides a consistent estimate of the variance of the causal effect when both models are correctly specified. In simulations we illustrate the utility of these results in difficult settings such as high dimensional covariate spaces or highly flexible models for the propensity score and outcome regression. Lastly, we analyze environmental exposure data from NHANES to identify how the effects of these exposures vary by subject level characteristics.

E0272: **Non-parametric causal effects based on longitudinal modified treatment policies**
Presenter: **Ivan Diaz**, Weill Cornell Medicine, United States

Most causal inference methods consider counterfactual variables under interventions that set the treatment deterministically. With continuous or multi-valued treatments or exposures, such counterfactuals may be of little practical interest because no feasible intervention can be implemented that would bring them about. Furthermore, violations to the positivity assumption, necessary for identification, are exacerbated with continuous and multi-valued treatments and deterministic interventions. We propose longitudinal modified treatment policies (LMTPs) as a non-parametric alternative. LMTPs can be designed to guarantee positivity, and yield effects of immediate practical relevance with an interpretation that is familiar to regular users of linear regression adjustment. We study the identification of the LMTP parameter, study properties of the statistical estimand such as the efficient influence function, and propose four different estimators. Two of our estimators are efficient, and one is sequentially doubly robust in the sense that it is consistent if, for each time point, either an outcome regression or a treatment mechanism is consistently estimated. We perform a simulation study to illustrate the properties of the estimators, and present the results of our motivating study on hypoxemia and mortality in Intensive Care Unit (ICU) patients. Software implementing our methods is provided in the form of the open source Rpackage lmtp freely available on GitHub.

E0273: **Deep learning for individual heterogeneity: An automatic inference framework**
Presenter: **Max Farrell**, University of Chicago, United States

A methodology is developed for estimation and inference using machine learning to enrich economic models. The framework takes a standard economic model and recasts the parameters as fully flexible nonparametric functions, to capture the rich heterogeneity based on potentially high dimensional or complex observable characteristics. These "parameter functions" retain all the interpretability, economic meaning, and discipline of classical parameters. We show that deep learning is well-suited to the structured modeling of heterogeneity in economics. First, we show how the network architecture can be easily designed to match the global structure of the economic model, delivering a novel methodology that moves deep learning away from prediction. Second, we prove convergence rates for the estimated parameter functions. These parameter functions are then the key input into the finite-dimensional parameter of inferential interest. We obtain valid inference based on a novel orthogonal score or influence function calculation that covers any second-stage parameter and any machine-learning-enriched model that uses a smooth per-observation loss function. No additional derivations are required, and the score can be taken directly to data, using automatic differentiation if needed to obtain the components. Our framework covers, as special cases, well-known examples such as average treatment effects and partially linear models, but we also seamlessly deliver new results.

EO160 Room K0.16 (Hybrid 02) BAYESIAN NONPARAMETRICS AND SEMIPARAMETRICS WITH APPLICATIONS Chair: Michael Daniels

E0391: **A Bayesian semi-parametric approach for inference on the PPCM from longitudinal data with dropout**
Presenter: **Maria Josefsson**, Centre for Demographic and Ageing Research, Sweden
Co-authors: Michael Daniels, Sara Pudas

Studies of memory trajectories using longitudinal data often result in highly non-representative samples due to selective study enrollment and attrition. An additional bias comes from practice effects that result in improved or maintained performance due to familiarity with test content or context. These challenges may bias study findings and severely distort the ability to generalize to the target population. We propose an approach for estimating the finite population mean of a longitudinal outcome conditioning on being alive at a specific time point. We develop a flexible Bayesian semi-parametric predictive estimator for population inference when longitudinal auxiliary information is known for the target population. We evaluate the sensitivity of the results to untestable assumptions and further compare our approach to other methods used for population inference in a simulation study. The proposed approach is motivated by 15-year longitudinal data from the Betula longitudinal cohort study. We apply our approach to estimating lifespan trajectories in episodic memory, with the aim to generalize findings to a target population.

E1017: **A prior based on allelic partitions for record linkage applications**
Presenter: **Brenda Betancourt**, University of Florida, United States

In database management, record linkage aims to identify multiple records that correspond to the same individual. This task can be treated as a clustering problem, in which a latent entity is associated with one or more noisy database records. However, in contrast to traditional clustering applications, a large number of clusters with a few observations per cluster is expected in this context. We introduce a new class of prior distributions based on allelic partitions that is especially suited for the small cluster setting of record linkage. We also introduce a set of novel microclustering conditions in order to impose further constraints on the cluster sizes a priori. We evaluate the performance of our proposed class of priors using simulated data and official statistics data sets, and show that our models provide competitive results compared to state-of-the-art microclustering models in the record linkage literature.

E0894: **Hierarchical Bayesian bootstrap for heterogeneous treatment effect estimation**
Presenter: **Arman Oganisian**, Brown University, United States
Co-authors: Nandita Mitra, Jason Roy

A major focus of causal inference is the estimation of heterogeneous average treatment effects (HTE) - average treatment effects within strata of another variable of interest such as levels of a biomarker, education, or age strata. Inference involves estimating a stratum-specific regression and integrating it over the distribution of confounders in that stratum - which itself must be estimated. Standard practice involves estimating these stratum-specific confounder distributions independently (e.g. via the empirical distribution or Bayesian bootstrap), which becomes problematic for sparsely populated strata with few observed confounder vectors. We develop a hierarchical Bayesian bootstrap (HBB) prior that induces a dependence across the stratum-specific confounder distributions. The HBB partially pools the stratum-specific distributions, allowing principled borrowing of confounder information across strata when sparsity is a concern. We show that posterior inference under the HBB can yield efficiency gains over standard marginalization approaches while avoiding strong parametric assumptions about the confounder distribution.

E1121: **In nonparametric and high-dimensional models, ignorability is an informative prior**
Presenter: **Antonio Linero**, University of Texas at Austin, United States

In problems with substantial missing data one in general must model two distinct data generating processes: the outcome process which generates the outcome and the missing data mechanism which determines the outcomes that we observe. Under the ignorability assumption, however, likelihood-based inference for the outcome process does not depend on the missing data mechanism so that only the outcome process needs to be modeled; because of this simplification, ignorability is often used as a baseline assumption. We study the implications of ignorability in the Bayesian context when there are high-dimensional nuisance parameters. We argue that ignorability is typically incompatible with sensible prior beliefs about the degree of selection bias, and show that for many problems ignorability directly implies that the selection bias is small with high prior probability. As examples, we consider semiparametric regression with Gaussian processes, high-dimensional ridge regression, and spike-and-slab priors.

EO479 Room K0.18 (Hybrid 03) RECENT ADVANCES IN CHANGE POINT ANALYSIS Chair: George Michailidis

E1589: **Inference for change points in high dimensional mean shift models**
Presenter: **Abhishek Kaul**, Washington State University, United States
Co-authors: George Michailidis

The problem of constructing confidence intervals for the locations of changepoints in a high-dimensional mean-shift model is considered. To that end, we develop a locally refitted least squares estimator and obtain component-wise and simultaneous rates of estimation of the underlying change points. The simultaneous rate is the sharpest available in the literature by at least a factor of log p, while the component-wise one is optimal. These results enable the existence of limiting distributions. Component-wise distributions are characterized under both vanishing and non-vanishing jump size regimes, while joint distributions for any finite subset of change point estimates are characterized under the latter regime, which also yields asymptotic independence of these estimates. The combined results are used to construct asymptotically valid component-wise and simultaneous confidence intervals for the changepoint parameters. The results are established under a high dimensional scaling, allowing for diminishing jump sizes, in the presence of a diverging number of change points and under subexponential errors. They are illustrated on synthetic data and on sensor measurements from smartphones for activity recognition.

E1667: **Optimistic search strategy in change point detection for large-scale data**
Presenter: **Housen Li**, University of Goettingen, Germany

As a classical and ever reviving topic, change point detection is often formulated as a search for the maximum of a gain function describing improved fits when segmenting the data. Searching through all candidate split points on the grid for finding the best one requires $O(T)$ evaluations of the gain function for an interval with T observations. If each evaluation is computationally demanding (e.g. in high-dimensional models), this can become infeasible. Instead, we propose optimistic search strategies with $O(logT)$ evaluations exploiting the specific structure of the gain function. Towards solid understanding of our strategies, we investigate in detail the Gaussian change in mean setup. For some of our proposals, we prove asymptotic minimax optimality for single and multiple change point scenarios. Our search strategies generalize far beyond the theoretically analyzed setup. We illustrate, as an example, the massive computational speedup in change point detection for high-dimensional Gaussian graphical models. More generally, we demonstrate empirically that optimistic search methods lead to competitive estimation performance while heavily reducing run-time.

E1697: **Inference for location of change points in high-dimensional non-stationary vector auto-regressive models**
Presenter: **Abolfazl Safikhani**, University of Florida, United States

Piece-wise stationary Vector Auto-Regressive models (VAR) are among the well-known and useful models in time series analysis. Existing methods provide sub-optimal estimators to detect the location of change/breakpoints in high-dimensional VAR models due to the existence of terms such as total sparsity of transition matrices and the logarithm of the number of time series components in the consistency rate. We study a refitted least squares estimator for change point parameters in high-dimensional VAR models with sparse model parameters. We show that the newly defined estimator reaches an optimal rate of convergence and the corresponding rate for relative location of change points reaches $O(1/T)$ for certain non-vanishing jump sizes, where T is the sample size. Further, the limiting distribution of the proposed estimate is obtained under both vanishing and non-vanishing jump sizes, thereby allowing the construction of confidence intervals for change point parameters. The proposed methodology is tested empirically over different synthetic data sets while an application to analyzing an EEG data set is also provided.

E1700: **Multiple change point detection in high dimensional data streams via the doubling algorithm**
Presenter: **George Michailidis**, University of Florida, United States

The problem of change point detection in high dimensional data has received a lot of attention in the literature, due to numerous applications in engineering, health and social sciences. A number of algorithms have been proposed and their theoretical properties investigated and performance in synthetic and real data sets illustrated. However, in many settings, the presence of a change point is driven by changes across a large number of the data streams under consideration. It is then reasonable to assume that detection of the locations of the underlying change points can be accomplished by examining only a subset of the available data streams, thus leading to significant computational gains. To that end, an algorithm based on binary segmentation is introduced that selects subsets of a certain size of the data streams and compares their detected change points; if there is an agreement, the algorithm stops, otherwise, it selects a fresh subset of data streams, double the sizes of the original ones, and continues in the same fashion until the agreement is reached or all the streams are included. Theoretical properties of the developed doubling algorithm are established and its performance is illustrated through numerical experiments.

EO489 Room K0.19 (Hybrid 04) RECENT ADVANCES IN COPULA METHODS (VIRTUAL) Chair: Radu Craiu

E0295: **Approximate Bayesian conditional copulas**
Presenter: **Clara Grazian**, University of New South Wales, Australia
Co-authors: Luciana Dalla Valle, Brunero Liseo

Copula models are flexible tools to represent complex structures of dependence for multivariate random variables. According to Sklar's theorem, any d-dimensional absolutely continuous distribution function can be uniquely represented as a copula, i.e. a joint cumulative distribution function on the unit hypercube with uniform marginals, which captures the dependence structure among the vector components. In real data applications, the interest of the analyses often lies on specific functionals of the dependence, which quantify aspects of it in a few numerical values. A broad literature exists on such functionals; however, extensions to include covariates are still limited. This is mainly due to the lack of unbiased estimators of the conditional copula, especially when one does not have enough information to select the copula model. We will present and compare several Bayesian methods to approximate the posterior distribution of functionals of the dependence varying according to covariates; the main advantage of the methods investigated here is that they use nonparametric models, avoiding the selection of the copula, which is usually a delicate aspect of copula modelling. These methods are compared in simulation studies and in two realistic applications, from civil engineering and astrophysics.

E0303: **Test of serial dependence for multivariate time series with arbitrary distributions**
Presenter: **Bouchra Nasri**, U. Montraal, Canada

Tests of serial independence are presented for a fixed number of consecutive observations from a stationary time series, first in the univariate case, and then in the multivariate case, where even vectors of large dimensions can be used. The common distribution function of the time series is not

assumed to be continuous, and the tests statistics are based on the multilinear copula process. A case study using a time series of Arctic sea ice extent images is used to illustrate the usefulness of the methodologies presented.

E1222: **Statistics of Wasserstein distributionally robust estimators**
Presenter: **Jose Blanchet**, Stanford, United States

Copulas provide an approach for estimating and characterizing joint distributions. The Wasserstein distance constructs the minimum cost copula (according to a specified criterion) between two marginal distributions. In recent years, a paradigm for robust estimation has emerged based on using this type of copula construction. Given a loss function to be minimized based on an empirical sample, a Wasserstein distributional robust estimator is obtained by choosing a parameter to minimize an expected loss against an adversary (say nature) that wishes to maximize the loss by choosing an appropriate probability model which is coupled with the data according to the Wasserstein distance given a budget constraint. It turns out that by appropriately choosing the loss and the geometry of the Wasserstein distance one can recover many classical statistical estimators (including Lasso, Graphical Lasso, SVM, group Lasso, among many others). A wide range of rich statistical quantities associated with these formulations is studied; for example, the optimal (in a certain sense) choice of the adversarial perturbation, weak convergence of natural confidence regions associated with these formulations, and asymptotic normality of the DRO estimators.

E1279: **Copula modelling of serially correlated multivariate data with hidden structures**
Presenter: **Robert Zimmerman**, University of Toronto, Canada
Co-authors: Vianey Leos Barajas, Radu Craiu

In applications where streams of data exhibit variable latent structures, it is natural to model the data-generating process as a finite-state hidden Markov model (HMM). When observing vectorial outcomes, we consider multivariate state-dependent distributions that are fused together by copulas. Such a "copula-within-HMM" framework is highly flexible, because it provides the freedom to vary both the marginal distributions of observed outcomes and the copula that determines the dependencies between them. However, inference for this model is not straightforward; while the EM algorithm is the standard technique for parameter estimation within HMMs, a direct application becomes unwieldy in the face of the additional model complexity brought about by the copula. We develop a robust and efficient EM algorithm for the copula-within-HMM model, and show that it performs well in both model estimation and state classification tasks on a variety of simulated and real-world datasets.

EO148 Room K0.20 (Hybrid 05) LAST TRENDS IN CLUSTERING AND CLASSIFICATION METHODS Chair: Marta Nai Ruscone

E0403: **Combining user-based collaborative filtering and classification for matching footwear size**
Presenter: **Irene Epifanio**, Universitat Jaume I, Spain
Co-authors: Aleix Alcacer, Jorge Valero, Alfredo Ballester

Size mismatch is a serious problem in online footwear purchases because size mismatch implies an almost sure return. Not only foot measurements are important in selecting a size, but also user preference. Therefore, we propose several methodologies that combine the information provided by a classifier with anthropometric measurements and user preference information through user-based collaborative filtering. As novelties: (1) the information sources are 3D foot measurements from a low-cost 3D foot digitizer, past purchases and self-reported size; (2) we propose to use an ordinal classifier after imputing missing data with different options based on the use of collaborative filtering; (3) we also propose an ensemble of ordinal classification and collaborative filtering results; and (4) several methodologies based on clustering and archetype analysis are introduced as user-based collaborative filtering for the first time. The hybrid methodologies were tested in a simulation study, and they were also applied to a dataset of Spanish footwear users. The results show that combining the information from both sources predicts the foot size better, and the new proposals provide better accuracy than the classic alternatives considered.

E0663: **How many data clusters are in the Galaxy data set: Bayesian cluster analysis in action**
Presenter: **Bettina Gruen**, WU (Vienna University of Economics and Business), Austria
Co-authors: Gertraud Malsiner-Walli, Sylvia Fruehwirth-Schnatter

In model-based clustering, the Galaxy data set is often used as a benchmark data set to study the performance of different modeling approaches. Based on results reported for the Galaxy data set using different Bayesian approaches, concerns were raised because the prior assumptions imposed remained rather obscure while playing a major role in the results obtained and conclusions drawn. We address these concerns by shedding light on how the specified priors influence the number of estimated clusters. We perform a sensitivity analysis of different prior specifications for the mixtures of a finite mixture model, i.e., the mixture model where a prior on the number of components is included. We use an extensive set of different prior specifications in a full factorial design and assess their impact on the estimated number of clusters for the Galaxy data set. Results highlight the interaction effects of the prior specifications and provide insights into which prior specifications are recommended to obtain a sparse clustering solution. A simulation study with artificial data provides further empirical evidence to support the recommendations. A clear understanding of the impact of the prior specifications removes restraints preventing the use of Bayesian methods due to the complexity of selecting suitable priors.

E0906: **On clustering and outlier detection with missing data**
Presenter: **Cristina Tortora**, San Jose State University, United States
Co-authors: Hung Tong , Louis Tran

Cluster analysis is a data analysis technique that aims to produce smaller groups of similar observations in a data set. In model-based clustering, the population is assumed to be a convex combination of sub-populations, each of which is modeled by a probability distribution. When the data are characterized by outliers the multivariate Student-t (T) and the contaminated normal distribution (CN) provide robust parameter estimates and therefore are more suitable choices compared to Gaussian Mixture models. Recently, the T and CN distributions have been extended to accommodate different tail behaviors across principal components, the models are referred to as multiple scaled distributions, i.e., MST and MSCN respectively. The mixture of CN has the advantage of automatically detecting outliers while the MSCN distribution, has the advantage of directional robust parameter estimates and outlier detection. The term "directional" implies that the parameter estimation and outlier detection procedures work separately for each principal component. Some practical limitations of the mentioned models are that they require the number of clusters to be known and the data set to be complete. The two mentioned limitations are overcome by providing a study of indices to select the number of clusters and presenting recent extensions of the CN and MSCN mixtures to cluster data that contain values missing at random.

E0812: **Simultaneous semi-parametric estimation of clustering and regression**
Presenter: **Matthieu Marbac**, CREST - ENSAI, France
Co-authors: Mohammed Sedki, Christophe Biernacki, Vincent Vandewalle

The parameter estimation of regression models with fixed group effects is investigated when the group variable is missing while group-related variables are available. This problem involves clustering to infer the missing group variable based on the group-related variables, and regression to build a model on the target variable given the group and eventually some additional variables. Thus, this problem can be formulated as the joint distribution modeling of the target and of the group-related variables. The usual parameter estimation strategy for this joint model is a two-step approach starting by learning the group variable (clustering step) and then plugging in its estimator for fitting the regression model (regression step). However, this approach is suboptimal (providing, in particular, biased regression estimates) since it does not make use of the target variable for clustering. Thus, we advise the use of a simultaneous estimation approach of both clustering and regression, in a semi-parametric framework. Numerical experiments illustrate the benefits of our proposition by considering wide ranges of distributions and regression models. The relevance

of our new method is illustrated in real data dealing with problems associated with high blood pressure prevention. The proposed approach is implemented in the R package ClusPred available on CRAN.

EO816 Room Virtual R21 DEVELOPMENTS IN OUTPUT ANALYSIS FOR MARKOV CHAIN MONTE CARLO Chair: James Flegal

E0514: **Beyond optimal online variance estimation in time series**
Presenter: **Kin Wai Chan**, The Chinese University of Hong Kong, Hong Kong
Co-authors: Man Fung Leung

The long-run variance (LRV) is an important quantity in the inference of dependent data. Recent advances in stochastic approximation show that online estimates of the LRV can be used to further improve learning algorithms. Nevertheless, existing 'optimal' LRV estimators face an efficiency dilemma and do not align with practical interests. We develop a general framework that uniformly improves and accelerates any LRV estimators. The main contributions lie in three aspects. Statistically, we propose several principles that lead to an online estimator with super-optimal statistical efficiency as compared with the offline counterpart. We also derive the first sufficient condition for a general estimator to be updated in $O(1)$ time or space. Computationally, we introduce mini-batch estimation to accelerate any online estimators in practice. Implementation issues such as automatic optimal parameters selection and multivariate extension are discussed. Practically, we apply our estimators to convergence diagnostics in Markov chain Monte Carlo methods and learning rate tuning in stochastic gradient methods. Our experiments show that the finite-sample properties of our proposals are in line with the theoretical findings.

E0681: **Optimal thinning of MCMC output**
Presenter: **Marina Riabiz**, King's College London, United Kingdom
Co-authors: Wilson Ye Chen, Jon Cockayne, Pawel Swietach, Steven Niederer, Chris Oates

The use of heuristics to assess the convergence and compress the output of Markov chain Monte Carlo can be sub-optimal in terms of the empirical approximations that are produced. Typically a number of the initial states are attributed to burn in and removed, whilst the remainder of the chain is thinned if compression is also required. We consider the problem of retrospectively selecting a subset of states, of fixed cardinality, from the sample path such that the approximation provided by their empirical distribution is close to optimal. A novel method is proposed, based on greedy minimisation of a kernel Stein discrepancy, that is suitable when the gradient of the log-target can be evaluated and an approximation using a small number of states is required. Theoretical results guarantee consistency of the method and its effectiveness is demonstrated in the challenging context of parameter inference for ordinary differential equations. Software is available in the Stein Thinning package in Python, R and MATLAB at http://stein-thinning.org/.

E0790: **Convergence of position-dependent MALA with application to conditional simulation in GLMMs**
Presenter: **Vivekananda Roy**, Iowa State University, United States

After discussing different variants of the Metropolis adjusted Langevin algorithms (MALA), we describe some convergence results for these Markov chains. The likelihood function in generalized linear mixed models (GLMMs) is available only as a high dimensional integral, and thus the resulting posterior densities in GLMMs are intractable. We study and compare the performance of variants of MALA in the context of conditional simulation from the two most popular GLMMs, namely the binomial GLMM with logit link and the Poisson GLMM with log link.

E0816: **Lugsail lag windows for estimating time-average covariance matrices**
Presenter: **James Flegal**, University of California - Riverside, United States
Co-authors: Dootika Vats

Lag windows are commonly used in time series, econometrics, steady-state simulation, and Markov chain Monte Carlo to estimate time-average covariance matrices. In the presence of a positive correlation of the underlying process, estimators of this matrix almost always exhibit significant negative bias, leading to undesirable finite-sample properties. We propose a new family of 15 lag windows specifically designed to improve finite-sample performance by offsetting this negative bias. Any existing lag window can be adapted into a lugsail equivalent with no additional assumptions. We use these lag windows within spectral variance estimators and demonstrate their advantages in a linear regression model with autocorrelated and heteroskedastic residuals. We further employ the lugsail lag windows in weighted batch means estimators due to their computational efficiency on large simulation output. We obtain bias and variance results for these multivariate estimators and significantly weaken the mixing condition on the process. Superior finite-sample properties are illustrated in a vector autoregressive process and a Bayesian logistic regression model.

EO116 Room Virtual R22 ESTIMATION AND INFERENCE FOR PRECISION MEDICINE Chair: Erica Moodie

E0462: **Estimation and inference on high-dimensional individualized treatment rule in observational data**
Presenter: **Yingqi Zhao**, Fred Hutchinson Cancer Research Center, United States

With the increasing adoption of electronic health records, there is an increasing interest in developing 25 individualized treatment rules, which recommend treatments according to patients characteristics, from large observational data. However, there is a lack of valid inference procedures for such rules developed from this type of data in the presence of high-dimensional covariates. We develop a penalized doubly robust method to estimate the optimal individualized treatment rule from high-dimensional data. We propose a split-and-pooled de-correlated score to construct hypothesis tests and confidence intervals. 30 Our proposal utilizes data splitting to conquer the slow convergence rate of nuisance parameter estimations, such as non-parametric methods for outcome regression or propensity models. We establish the limiting distributions of the split-and-pooled de-correlated score test and the corresponding one-step estimator in the high-dimensional setting. Simulation and real data analysis are conducted to demonstrate the superiority of the proposed method.

E0587: **Estimation of optimal treatment regimes with censored time-to-event outcome: A classification perspective**
Presenter: **Marie Davidian**, North CArolina State University, United States

Clinicians make a series of decisions at key points over the course of a patient's disease or disorder based on a synthesis of evolving information on the patient. A treatment regime is a sequence of decision rules mapping a patient's history to the set of feasible treatment options and thus formalizes this process. An optimal regime is one leading to the most beneficial outcome on average if used to make treatment decisions for the patient population. In many chronic disease contexts, the outcome is a possibly censored time to an event. We describe a method for estimation of an optimal regime with such outcomes, which is based on maximizing a locally efficient, doubly robust, augmented inverse probability weighted estimator for average outcome over a class of regimes. This optimization can be cast as a classification problem, which allows well-known methodology for classification to be exploited in a backward iterative algorithm. The performance of the method is demonstrated.

E1065: **Bayesian semiparametric approaches to tailoring strategies**
Presenter: **David Stephens**, McGill University, Canada

Frequentist semiparametric methods for optimizing and tailoring treatments to patient characteristics are well established, but typically rely on asymptotic justifications, for example, sandwich estimation, for uncertainty representations. We will present computational Bayesian methods that give exact posterior credible intervals (under an assumed flexible semiparametric specification) in the treatment-tailoring problem. These Monte Carlo-based approaches rely on a Bayesian nonparametric formulation, and can be applied in regression-based and inverse weighting approaches.

E1091: **Post-contextual-bandit inference**
Presenter: **Nathan Kallus**, Cornell University, United States

Contextual bandit algorithms are increasingly replacing non-adaptive A/B tests in e-commerce, healthcare, and policymaking because they can both improve outcomes for study participants and increase the chance of identifying good or even best policies. To support credible inference on novel interventions at the end of the study, nonetheless, we still want to construct valid confidence intervals on average treatment effects, subgroup effects, or value of new policies. The adaptive nature of the data collected by contextual bandit algorithms, however, makes this difficult: standard estimators are no longer asymptotically normally distributed and classic confidence intervals fail to provide correct coverage. While this has been addressed in non-contextual settings by using stabilized estimators, the contextual setting poses unique challenges that we tackle for the first time. We propose the Contextual Adaptive Doubly Robust (CADR) estimator, the first estimator for policy value that is asymptotically normal under contextual adaptive data collection. The main technical challenge in constructing CADR is designing adaptive and consistent conditional standard deviation estimators for stabilization. Extensive numerical experiments using 57 OpenML datasets demonstrate that confidence intervals based on CADR uniquely provide correct coverage.

EO086 Room Virtual R23 RECENT ADVANCES IN OPTIMAL EXPERIMENTAL DESIGN Chair: Saumen Mandal

E1560: **Circuit bases in optimal experimental design and randomization**
Presenter: **Henry Wynn**, London School of Economics, United Kingdom
Co-authors: Fabio Rapallo, Roberto Fontana

Circuits are a construction used extensively in integer programming, numerical linear algebra and the application of toric ideals in algebraic statistics. Recent work is brought together. Circuits have been previously used to construct small sample optimally robust factorial designs and to construct randomisation schemes to de-bias the analysis of factorial experiments. The Binet-Cauchy theorem has been used alongside the circuit theory to construct robust designs. A subset of binary (0-1) circuits turned out to provide the set of generators, under set partition theory, for all valid randomization schemes. In both cases, the circuits are for the kernel K of the usual X-matrix of the model/design pair, namely the residual space. At least one connection is that the extent of the available randomisation schemes depends on the extent of symmetry in the design, which is also a feature of optimal design. Heuristically: good designs lead to versatility in randomization. The complexity of the circuit structure even for small problems requires computer algebra. The package 4ti2 is suggested. The methods are applied to some standard and new examples, revealing deep combinatorial structures.

E1725: **Computing optimal regression designs with multiple objectives**
Presenter: **Julie Zhou**, University of Victoria, Canada

Model-based optimal regression designs with multiple objectives are common in practice. The objectives are often competitive. It is extremely hard to derive analytical solutions for optimal designs with multiple objectives, and there are also no general and efficient algorithms for searching such designs for user-specified nonlinear models and criteria. We propose a new and effective approach for finding multiple-objective optimal designs via the CVX solver. It can efficiently find different types of multiple-objective optimal designs after the optimization problems are carefully formulated as convex optimization problems. This approach is flexible and can be applied to any regression model. We present several applications including minimax and efficiency constrained multiple-objective optimal designs.

E1616: **Bayesian optimal designs with high prediction efficiency**
Presenter: **Po Yang**, University of Manitoba, Canada

Design of experiments is a strategy used to identify the important factors which affect the response. A well-designed experiment plays a vital role in industry since it can provide information to conduct time- and cost-efficient processes. For response surface experiments, the prediction of the response is an important task. We propose Bayesian optimality criteria for constructing optimal designs that have high prediction efficiency and less dependence on an assumed model. The constructed designs are compared with the designs obtained using different optimality criteria.

E1704: **Binary response models comparison using the alpha-Chernoff divergence measure and exponential integral functions**
Presenter: **Subir Ghosh**, University of California, United States

The families of binary response models describe the data on a response variable having two possible outcomes and explanatory variables when the possible responses and their probabilities are functions of the explanatory variables. The alpha-Chernoff divergence and the Bhattacharyya divergence measures when alpha equals one-half are the criterion functions used for quantifying the dissimilarity between probability distributions by expressing the divergence measures in terms of the exponential integral functions. In addition, the dependences of odds ratio and hazard function on the explanatory variables are also a part of the modeling.

EO607 Room Virtual R24 ADVANCE STATISTICAL TOOLS FOR MODERN HIGH DIMENSIONAL DATA Chair: Shahina Rahman

E1358: **Vector-valued infinite task learning in style transfer**
Presenter: **Zoltan Szabo**, LSE, United Kingdom

Style transfer is a central problem of machine learning. In various applications of style transfer, however, there is a continuum of styles to handle. We show how one can leverage vector-valued reproducing kernel Hilbert spaces and infinite task learning to tackle this challenge in a principled way. The approach is instantiated in emotion transfer, achieving low reconstruction cost on various benchmarks.

E1418: **Solution path based variable selection**
Presenter: **Karl Gregory**, University of South Carolina, United States

Methods are considered for variable selection in high-dimensional regression based on the solution paths of sparse estimators. In particular, we measure the importance of a covariate by the amount of sparsity penalization it is able to overcome in order to enter the model, for example under LASSO penalization or along the path of the least angle regression algorithm. In addition, we consider measuring the importance of a covariate by how much the solution path of a sparsity-promoting estimator changes when the covariate is removed from the model. We study the performance of variable screening procedures based on these metrics as well as the performance of bootstrap methods for estimating their distributions under certain null hypotheses in the regression coefficients.

E1438: **Long-term prediction for high-dimensional regression**
Presenter: **Sayar Karmakar**, University of Florida, United States

Time-aggregated prediction intervals are constructed for a univariate response time series in a high-dimensional regression regime. A simple quantile-based approach on the LASSO residuals seems to provide reasonably good prediction intervals. We allow for a very general possibly heavy-tailed, possibly long-memory and possibly non-linear dependent error process and discuss both the situations where the predictors are assumed to form a fixed or stochastic design. Finally, we construct prediction intervals for hourly electricity prices over horizons spanning 17 weeks and compare them to selected Bayesian and bootstrap interval forecasts

E1578: **A fast and automated clustering of large scale high dimensional data: Application to scRNA-seq data**
Presenter: **Shahina Rahman**, Texas A&M University, United States
Co-authors: Suhasini Subbarao, Valen Johnson

Technological advancements are now occurring at a breathtaking speed, thus allowing researchers to collect a massive volume of data. For example,

in biomedical engineering, using next-generation sequencing technologies, it is now possible to profile the transcriptome of individual cells. This technology can provide detailed catalogs of millions of cells found in a sample. Despite the availability of a large number of clustering algorithms, very few of them can be applied to these massive high dimensional datasets due to large computational costs and the lack of reliability of their results. To address such issues, we have developed a fast and scalable clustering algorithm based on the Gram matrix transformation. The major advantage of this clustering method over other competitors is that it lacks major tuning parameters and runs in linear time. Besides, under mild assumptions, the method also provides a theoretical guarantee on its result.

EO384 Room Virtual R25 GEOMETRY AND TOPOLOGY IN STATISTICS AND MACHINE LEARNING Chair: Wolfgang Polonik

E0640: **Fractal Gaussian networks: A sparse random graph model based on gaussian multiplicative chaos**
Presenter: **Krishnakumar Balasubramanian**, University of California, Davis, United States

A novel stochastic network model is introduced, which is called Fractal Gaussian Network (FGN), that embodies well-defined and analytically tractable fractal structures. FGNs are driven by the latent spatial geometry of Gaussian Multiplicative Chaos (GMC), a canonical model of fractality in its own right. They interpolate continuously between the popular purely random geometric graphs (aka the Poisson Boolean network), and random graphs with increasingly fractal behavior. After introducing and motivating the model, I will discuss some probabilistic (e.g., expected edge/triangle counts, spectral properties) and statistical question (e.g, detecting the presence of fractality and parameter estimation based on observed network data) related to FGNs.

E1034: **Confidence regions for filamentary structures**
Presenter: **Wanli Qiao**, George Mason University, United States

Filamentary structures, also called ridges, generalize the concept of modes of density functions and provide low-dimensional representations of point clouds. Using kernel type plug-in estimators, we give asymptotic confidence regions for filamentary structures based on two bootstrap approaches: multiplier bootstrap and empirical bootstrap. Our theoretical framework differs from many existing ones in the literature in that we allow the possible existence of intersections to emphasize the nontrivial topology of filamentary structures. Our confidence regions are asymptotically valid in different scenarios depending on how flat the filamentary structures are.

E1479: **Modeling shapes and fields**
Presenter: **Sayan Mukherjee**, Duke University, United States

Modeling shapes and fields is considered via topological and lifted-topological transforms. Specifically, we show how the Euler Characteristic Transform and the Lifted Euler Characteristic Transform can be used in practice for statistical analysis of shape and field data. The Lifted Euler Characteristic is an alternative to the Euler calculus for real-valued functions. We also state a moduli space of shapes for which we can provide a complexity metric for the shapes. Furthermore, we provide a sheaf theoretic construction of shape space that does not require diffeomorphisms or correspondence. A direct result of this sheaf theoretic construction is that in three dimensions for meshes, 0-dimensional homology is enough to characterize the shape.

E1674: **On using graph distances to estimate euclidean and related distances**
Presenter: **Ery Arias-Castro**, UC San Diego, United States

Graph distances have proven quite useful in machine learning/statistics, particularly in the estimation of Euclidean or geodesic distances, and as such have been used to embed a graph (the multidimensional scaling problem). A partial review of the literature will be included and then more recent developments will be presented, including the minimax estimation of distances on a surface and consequences for manifold learning; the estimation of curvature-constrained distances on a surface; and the estimation of Euclidean distances based on an unweighted and noisy neighborhood graph.

EO244 Room Virtual R26 MODERN ADVANCED STATISTICAL METHODS IN BIOMEDICAL RESEARCH Chair: Esra Kurum

E0616: **Multilevel joint modeling of hospitalization and survival in patients on dialysis**
Presenter: **Esra Kurum**, University of California, Riverside, United States
Co-authors: Danh Nguyen, Damla Senturk

More than 720,000 patients with end-stage renal disease in the U.S. require life-sustaining dialysis treatment that is predominantly administered at local dialysis facilities. In this population of typically older patients with a high morbidity burden, hospitalization is frequent at a rate of about twice per patient-year. Aside from frequent hospitalizations, which are a major source of death risk, overall mortality in dialysis patients is higher than in other comparable populations, including Medicare patients with cancer. Thus, understanding patient- and facility-level risk factors that jointly contribute to longitudinal hospitalizations and mortality is of interest. Towards this objective, we propose a novel methodology to jointly model hospitalization, a binary longitudinal outcome, and survival, based on multilevel data from the United States Renal Data System, with repeated observations over time nested in patients and patients nested in dialysis facilities. Motivated by the data structure, the proposed joint modeling approach includes multilevel random effects and multilevel covariates, at both the patient and facility levels. An approximate Expectation-Maximization (EM) algorithm is developed for estimation where fully exponential Laplace approximations are utilized to address computational challenges and standard error formulas for the estimated parameters are derived.

E1177: **Enhanced doubly robust procedure for causal inference**
Presenter: **Ao Yuan**, Georgetown University, United States
Co-authors: Anqi Yin, Ming Tan

The doubly robust estimator is a popular tool in causal inference, which provides double protection of unbiasdness. However, most existing methods for such an estimator use parametric models, and are not robust enough. We propose a semi-parametric model for this estimator, in which both the propensity score and outcome models are semi-parametric, with a non-parametric link function to enhance the robustness and parametric regression effects for easy interpretation. Simulation studies are conducted to evaluate the performance of the proposed method and compared with existing parametric and naive methods, showing a clear advantage of the proposed method. The method is then applied to analyze real clinical trial data.

E1218: **Identifying treatment-sensitive subgroups based on multiple covariates and longitudinal measurements**
Presenter: **Yingwei Peng**, Queenś University, Canada

Identifying a subgroup of patients who may be sensitive to a specific treatment is an important step towards personalized medicine. We consider the effect of a treatment on longitudinal measurements, which may be continuous or categorical, such as quality of life scores assessed over the duration of a clinical trial. We assume multiple baseline covariates, such as age and expression levels of genes, are available and propose a generalized single-index linear threshold model to simultaneously identify the treatment-sensitive subgroup and assess the treatment-by-subgroup interaction. Because the model involves an indicator function with unknown parameters, conventional procedures are difficult to apply for the inferences of parameters in the model. We define smoothed generalized estimating equations and propose an inference procedure based on these equations with an efficient spectral algorithm employed to find their solutions. The proposed procedure is evaluated through simulation studies and application to the analysis of data from a randomized clinical trial in advanced pancreatic cancer.

E1450: **Convolutional neural networks estimation via Lipschitz loss functions with applications to biomedical studies**
Presenter: **Ke Huang**, University of California, Riverside, United States

Co-authors: Shujie Ma

The statistical learning theory of a convolutional neural networks (CNNs) estimator obtained from empirical risk minimization with a Lipschitz loss function is investigated. Our framework can be applied to both regression and classification problems. The CNNs estimator is constructed from a network architecture of CNNs followed by two fully connected layers. We establish an explicit bound for both the approximation error and the estimation error. The results provide theoretical guidance for selecting the number of convolutional layers and the number of nodes on each feed-forward layer in the considered network structure. For the purpose of alleviating "the curse of dimensionality", we further assume that the target function is defined on a low-dimensional manifold, and develop non-asymptotic excess risk bounds for our estimator. We illustrate the performance of the proposed method through Monte Carlo simulation experiments, and apply it to biomedical data applications.

EO340 Room Virtual R27 ADVANCES IN LONGITUDINAL DATA ANALYSIS Chair: Sanjoy Sinha

E0341: **Robust analysis of longitudinal mixed binary and count responses**

Presenter: **Sanjoy Sinha**, Carleton University, Canada

A robust technique for jointly analyzing longitudinal binary and count responses will be discussed. The robust method developed in the framework of the generalized linear mixed models is designed to provide protections against potential outliers in the response variables and covariates. The finite-sample properties of the robust estimators will be discussed using Monte Carlo simulations. An application will be presented using actual longitudinal data from a health survey.

E0406: **D-optimal designs for ordered logit and ordered probit models**

Presenter: **Xiaojian Xu**, Brock University, Canada

Both ordered logit and ordered probit models often serve as appropriate frameworks for statistical analysis when ordinal responses are involved. Although statistical inferences based on these ordinal regression models have been studied extensively in the literature, very few developments, have been done for optimizing the designs of experiments in the context of either ordered logit or ordered probit regression. We discuss optimal designs, particularly for ordered logit and ordered probit regression. Maximum likelihood estimation and D-optimality are adopted. Our resulting D-optimal designs are presented along with a comparison study that indicates D-optimal designs outperform their competitors. We also address the dependency issue of a design on the unknown parameters by an attainable two-stage design process. Furthermore, any assumed model form may not be accurate and lead to low efficiencies. Therefore, we also construct robust designs under the consideration of possible model departures.

E0697: **Modelling big data to predict trajectories of repeated binary outcomes**

Presenter: **Tariqul Hasan**, University of New Brunswick, Canada

Co-authors: Rafiqul Chowdhury

In environmental, health, social and biological sciences, the amount of longitudinal or repeated data being captured and stored is increasing significantly, due to technological advances and lower cost of data acquisition. Adequate modeling of such big data is useful for cost reduction, time reduction, new product developments, and developing new strategies and optimum decision making. As there is limited literature available to analyze big data in longitudinal or repeated measures setup, there is a significant interest to develop a big data modeling approach for naturally correlated longitudinal data. We develop a joint modelling approach to predict the trajectory risks of a sequence of repeated outcomes of interest. Trajectory prediction from big data collected longitudinally requires a unique modeling approach to overcome additional levels of complexity. The proposed methodology will be compared with various machine learning algorithms such as the Decision Tree, Random Forest, Support Vector Machine, Neural Network, etc. The performance of the proposed model will also be demonstrated using a longitudinally collected Fine particulate matter (PM2.5) data set.

E0708: **Statistical methods for clustered longitudinal binary data**

Presenter: **Leilei Zeng**, University of Waterloo, Canada

In many settings in experimental research, clusters of subjects (e.g. families/schools/clinics) are randomly assigned to different interventions and each subject has repeated measurements over the study period. The resulting responses are then cross-sectionally correlated within clusters at a given assessment time, and longitudinally correlated within subjects over time. Methods for clustered longitudinal binary data are proposed based on transition models and generalized estimating equations. Efficiency gains for the marginal regression parameters are realized when the intra-cluster association is strong. Guidance for the design of randomized trials involving this kind of analysis is also provided. The formula for the number of clusters is derived based on the robust variance estimators from the transition models that account for the intra-cluster associations.

EO372 Room Virtual R28 ADVANCES IN NETWORK ANALYSIS AND CLUSTERING Chair: Anderson Ye Zhang

E1752: **Power enhancement and phase transitions for global testing of the mixed membership stochastic block model**

Presenter: **Tracy Ke**, Harvard University, United States

The mixed-membership stochastic block model (MMSBM) is a common model for social networks. We are interested in testing $K = 1$ versus $K > 1$, where K is the number of communities. We first study the degree-based chi-square test and the orthodox Signed Quadrilateral (oSQ) test. We reveal that these two test statistics target to estimate an order-2 polynomial and an order-4 polynomial of a signal matrix, respectively. We derive the asymptotic null distribution and power for both tests. Unfortunately, for each test, there exists a parameter regime where its power is unsatisfactory. Next, we propose a Power Enhancement (PE) test by combining these two test statistics. We show that the PE test statistic converges in law to a chi-square distribution with 2 degrees of freedom, and that it has a better power compared with both the degree-based chi-square test and the oSQ test. To assess the optimality of global testing, we adopt the phase transition framework. We consider a random-membership MMSBM and discover an explicit quantity that defines the Region of Possibility and the Region of Impossibility. A test is called optimally adaptive if it successfully distinguishes any alternative hypothesis in the Region of Impossibility from any null hypothesis. We show that the PE test is optimally adaptive, but the chi-square test and the oSQ test are not.

E0198: **Individual-centered partial information in social networks**

Presenter: **Xin Tong**, University of Southern California, United States

Most existing statistical network analysis literature assumes a global view of the network, under which community detection, testing, and other statistical procedures are developed. Yet, people frequently make decisions based on their partial understanding of the network information in the real world. As individuals barely know beyond friends' friends, we assume that an individual of interest knows all paths of length up to $L = 2$ that originate from her. As a result, this individual's perceived adjacency matrix B differs significantly from the usual adjacency matrix A based on the global information. The new individual-centered partial information framework sparks an array of interesting endeavors from theory to practice. Key general properties on the eigenvalues and eigenvectors of BE, a major term of B, are derived. These general results, coupled with the classic stochastic block model, lead to a new theory-backed spectral approach to detecting the community memberships based on an anchored individual's partial information. Real data analysis delivers interesting insights that cannot be obtained from global network analysis.

E1549: **Community detection on mixture multi-layer networks via regularized tensor decomposition**

Presenter: **Dong Xia**, Hong Kong University of Science and Technology, Hong Kong

The focus is on the problem of community detection in multi-layer networks, where pairs of nodes can be related in multiple modalities. We introduce a general framework, i.e., mixture multi-layer stochastic block model (MMSBM), which includes many earlier models as special cases.

We propose a tensor-based algorithm (TWIST) to reveal both global/local memberships of nodes, and memberships of layers. We show that the TWIST procedure can accurately detect the communities with small misclassification errors as the number of nodes and/or the number of layers increases. Numerical studies confirm our theoretical findings. To our best knowledge, this is the first systematic study on the mixture multi-layer networks using tensor decomposition. The method is applied to two real datasets: worldwide trading networks and malaria parasite genes networks, yielding new and interesting findings.

E0231: **Diffusion source identification on networks with statistical confidence**
Presenter: **Tianxi Li**, University of Virginia, United States
Co-authors: Quinlan Dawkins, Haifeng Xu

Diffusion source identification on networks is a problem of fundamental importance in a broad class of applications, including rumor controlling and virus identification. Though this problem has received significant recent attention, most studies have focused only on very restrictive settings and lack theoretical guarantees for more realistic networks. We introduce a statistical framework for the study of diffusion source identification and develop a confidence set inference approach inspired by hypothesis testing. Our method efficiently produces a small subset of nodes, which provably covers the source node with any pre-specified confidence level without restrictive assumptions on network structures. To our knowledge, this is the first diffusion source identification method with a practically useful theoretical guarantee on general networks. We demonstrate our approach via extensive synthetic experiments on well-known random network models and a mobility network between cities concerning the COVID-19 spreading.

EO206 Room Virtual R29 ROBUST CAUSAL INFERENCE Chair: Zijian Guo

E0471: **Causal aggregation: Estimation and inference of causal effects by constraint-based data fusion**
Presenter: **Dominik Rothenhaeusler**, Stanford University, United States
Co-authors: Jaime Gimenez

Randomized experiments are the gold standard for causal inference. In experiments, usually, one variable is manipulated, and its effect is measured on an outcome. However, practitioners may also be interested in the effect of simultaneous interventions on multiple covariates on a fixed target variable. We discuss a method that allows estimating the effect of joint interventions using data from different experiments in which only very few variables are manipulated. The proposed method allows combining data sets arising from randomized experiments and observational data sets for which IV assumptions or unconfoundedness hold. Compared to existing approaches, the approach is applicable in settings with very little knowledge about the graph. This makes the approach potentially more reliable in cases where the practitioner has limited background knowledge. We demonstrate the effectiveness of the proposed method on synthetic and semi-synthetic data.

E0645: **Causal inference for nonlinear outcome models with possibly invalid instrumental variables**
Presenter: **Sai Li**, Renmin University of China, China
Co-authors: Zijian Guo

Instrumental variable methods are widely used for inferring the causal effect of an exposure on an outcome when the observed relationship is potentially affected by unmeasured confounders. Existing instrumental variable methods for nonlinear outcome models require stringent identifiability conditions. We develop a robust causal inference framework for nonlinear outcome models, which relaxes the conventional identifiability conditions. We adopt a flexible semi-parametric potential outcome model and propose new identifiability conditions for identifying the model parameters and causal effects. We devise a novel three-step inference procedure for the conditional average treatment effect and establish the asymptotic normality of the proposed point estimator. We construct confidence intervals for the causal effect by the bootstrap method. The proposed method is demonstrated in a large set of simulation studies and is applied to study the causal effects of lipid levels on whether the glucose level is normal or high over a mice dataset.

E0615: **Agglomerative hierarchical clustering for selecting valid instrumental variables**
Presenter: **Xiaoran Liang**, University of Bristol, United Kingdom

An instrumental variable (IV) selection procedure is proposed that combines the agglomerative hierarchical clustering method and the Hansen-Sargan overidentification test for selecting valid instruments for IV estimation from a large set of candidate instruments. Some of the instruments may be invalid in the sense that they fail the exclusion restriction. We show that if the largest group of IVs is valid, our method can achieve oracle selection and estimation results. Compared to the previous IV selection methods, our method has the advantage that it can deal with the weak instruments problem effectively, and can be easily extended to settings where there are multiple endogenous regressors and heterogeneous treatment effects. We conduct Monte Carlo simulations to examine the performance of our method. The simulation results show that our method achieves oracle selection and estimation results in both single and multiple endogenous regressors settings in large samples when instruments are strong. The method works well, even when many instruments are weak, with single or multiple regressors. We apply our method to the estimation of the effects of immigration on wages in the US.

E1149: **Ultra-high dimensional learning of polygenic risk scores for mendelian randomization studies**
Presenter: **Linbo Wang**, University of Toronto, Canada
Co-authors: Xinyi Zhang, Stanislav Volgushev, Dehan Kong

Mendelian randomization (MR) is a method by which genetic variants are leveraged as instrumental variables (IV) to investigate causal relationships between modifiable exposure or risk factor and a clinically relevant outcome from observational data. To provide reliable causal evidence, one key step in MR analysis is to identify valid instruments among the collection of all candidate genetic variants. Current methods work well when the size of candidate instruments is moderate. However, for the identification in ultrahigh dimensions, normal in practice, empirical evidence suggests that existing procedures may miss many or even all valid instrumental variables, due to the inclusion of irrelevant variables which exhibit spurious correlation with the exposure in observational data. To overcome this challenge, we propose a novel approach to first remove irrelevant variables from the candidate set and then apply existing work to the remaining candidates to make valid causal inference. Theoretically, we proved that causal effect estimates from selected irrelevant variables are also centered around a single value but distinct from the true causal effect with high probability, which makes selected irrelevant variables and valid instruments separable. Simulation studies and data application further demonstrate that the proposed procedure outperforms existing methods under ultrahigh-dimensional settings.

EO529 Room Virtual R30 NEW DEVELOPMENTS ON DATA DEPTH AND ITS APPLICATIONS Chair: Sara Lopez Pintado

E0319: **The enlarged simplicial depth**
Presenter: **Alicia Nieto-Reyes**, Universidad de Cantabria, Spain
Co-authors: Giacomo Francisci, Claudio Agostinelli

Most multivariate statistical data depth functions vanish right outside the convex hull of the support of the distribution with respect to which the depth is computed. This results in a challenge when we aim to distinguish among different points that are outside of the convex hull of the distribution support, with respect to which the depth is calculated. We give the first two definitions that overcome this issue for the simplicial depth function. We also present some corresponding depth estimators that do not vanish right outside the convex hull of the data. The properties of the

definitions and of the respective estimators are illustrated, theoretically and by performing Monte Carlo simulations. Additionally, the importance of such proposals will be shown through the analysis of real data sets.

E1416: **A pseudo-metric between probability distributions based on depth-trimmed regions**
Presenter: **Guillaume Staerman**, Telecom Paris, Institut Polytechnique de Paris, France
Co-authors: Pavlo Mozharovskyi

The design of a metric between probability distributions is a longstanding problem motivated by numerous applications in Machine Learning. Focusing on continuous probability distributions on a Euclidean space, we introduce a novel pseudo-metric between probability distributions by leveraging the extension of univariate quantiles to multivariate spaces. Data depth is a nonparametric statistical tool that measures the centrality of any element x with respect to (w.r.t.) a probability distribution or a data set. It is a natural median-oriented extension of the cumulative distribution function (cdf) to the multivariate case. Thus, its upper-level sets—the depth-trimmed regions—give rise to a definition of multivariate quantiles. The new pseudo-metric relies on the average of the Hausdorff distance between the depth-based quantile regions w.r.t. each distribution. After discussing the properties of this pseudo-metric inherited from data depth, we provide conditions under which it defines a distance. Interestingly, the derived non-asymptotic bound shows that in contrast to the widely used Wasserstein distance, the proposed pseudo-metric does not suffer from the curse of dimensionality. Robustness, an appealing feature of this pseudo-metric, is studied through the finite sample breakdown point. Moreover, we propose an efficient approximation method with linear time complexity w.r.t. the size of the data set and its dimension.

E1382: **Statistical data depth and its applications to health sciences**
Presenter: **Sara Lopez Pintado**, Northeastern University, United States

Data depth was originally introduced for multivariate data as a powerful non-parametric tool for developing robust exploratory data analysis methods. It provides a way of measuring how representative an observation is within the distribution or sample and of ranking multivariate observations from center-outward. Based on these depth-rankings, robust estimators and outliers can be defined. Notions of depth have been extended to functional data in the last several decades. We develop different depth-based methods for general functional data, such as an envelope test for detecting and visualizing differences between groups of functions. We applied this method to longitudinal growth data, where the goal is to find differences between the growth pattern of normal versus premature low birth weight babies. We also introduce and establish the properties of the metric halfspace depth, an extension of the well-known Tukey's depth to object data in general metric spaces. The metric halfspace depth was applied to an Alzheimer's disease study, revealing group differences in the brain connectivity, modeled as covariance matrices, for subjects in different stages of dementia.

E1489: **Flexible integrated functional depths**
Presenter: **Germain Van Bever**, Universite de Namur, Belgium
Co-authors: Stanislav Nagy, Pauliina Ilmonen, Sami Helander, Lauri Viitasaari

A new class of functional depths is introduced. A generic member of this class is coined Jth order kth moment integrated depth. It is based on the distribution of the cross-sectional halfspace depth of a function in the marginal evaluations (in time) of the random process. Asymptotic properties of the proposed depths are provided: we show that they are uniformly consistent and satisfy an inequality related to the law of the iterated logarithm. Moreover, limiting distributions are derived under mild local regularity assumptions. The versatility displayed by the new class of depths makes them particularly amenable for capturing important features of functional distributions. This is illustrated in supervised learning, where we show that the corresponding maximum depth classifiers outperform classical competitors.

EO082 Room Virtual R31 NEW DIRECTIONS IN FUNCTIONAL AND HIGH-DIMENSIONAL DATA ANALYSIS Chair: Alessia Caponera

E0371: **High-dimensional MANOVA via bootstrapping and its application to functional and sparse count data**
Presenter: **Miles Lopes**, UC Davis, United States
Co-authors: Zhenhua Lin, Hans-Georg Mueller

A new approach is proposed for the problem of high-dimensional multivariate ANOVA via bootstrapping max statistics. The proposed method is suited to simultaneously test the equality of several pairs of mean vectors of potentially more than two populations. By exploiting the "variance decay" property that is a natural feature in relevant applications, we are able to establish a dimension-free and nearly parametric convergence rate for bootstrap approximation. In addition, we illustrate the proposed approach with ANOVA problems for functional data and sparse count data. The proposed method is shown to work well in simulations and several real data applications.

E0633: **Multilevel hybrid principal components analysis forregion-referenced functional EEG data**
Presenter: **Damla Senturk**, University of California Los Angeles, United States

Electroencephalography (EEG) experiments produce region-referenced functional data representing brain signals in the time or the frequency domain collected across the scalp. The data typically also have a multilevel structure with high dimensional observations collected across multiple experimental conditions or visits. Common analysis approaches reduce the data complexity by collapsing the functional and regional dimensions, where event-related potential (ERP) features or band power are targeted in a pre-specified scalp region. This practice can fail to portray more comprehensive differences in the entire ERP signal or the power spectral density (PSD) across the scalp. Building on the weak separability of the high-dimensional covariance process, the proposed multilevel hybrid principal components analysis (M-HPCA) utilizes dimension reduction tools from both vector and functional principal components analysis to decompose the total variation into between- and within-subject variance. The resulting model components are estimated in a mixed-effects modeling framework via a computationally efficient minorization-maximization (MM) algorithm coupled with bootstrap. The diverse array of applications of M-HPCA is showcased with two studies of individuals with autism. While ERP responses to match vs. mismatch conditions are compared in an audio odd-ball paradigm in the first study, short-term reliability of the PSD across visits is compared in the second.

E0637: **Functional principal component analysis of cointegrated functional time series**
Presenter: **Won-Ki Seo**, University of Sydney, Australia

Functional principal component analysis (FPCA) has played an important role in the development of functional time series analysis. The aim is to investigate how FPCA can be used to analyze cointegrated functional time series and proposes a modification of FPCA as a novel statistical tool. The modified FPCA not only provides an asymptotically more efficient estimator of the cointegrating vectors, but also leads to novel FPCA-based tests for examining some essential properties of cointegrated functional time series. As an empirical illustration, our methodology is applied to two empirical examples: U.S. age-specific employment rates and earning densities.

E0290: **Functional data analysis in constrained spaces**
Presenter: **John Aston**, University of Cambridge, United Kingdom
Co-authors: Eardi Lila

Traditional Functional Data has been analysed as 1-dimensional curves which are assumed to be i.i.d. However, in many settings, this is not the case. We will look at examples of functional data which is constrained to lie in certain spaces, which are often non-Euclidean. This limits usual estimation techniques and requires the development of new methods to take into account the geometry of the setting. We will illustrate the issues with data from neuroimaging which is both surface-based and is related to the covariance operator of a process.

EO689 Room Virtual R32 ADVANCES IN CLUSTERING, NETWORK ANALYSIS, AND MULTIVARIATE STATISTICS Chair: Joshua Cape

E1370: **Fast and interpretable consensus clustering via minipatch learning**
Presenter: **Genevera Allen**, Rice University, United States

Consensus clustering has been widely used in bioinformatics and other applications to improve the accuracy, stability and reliability of clustering results. This approach ensembles cluster co-occurrences from multiple clustering runs on subsampled observations. For application to large-scale bioinofmratics data, consensus clustering has two significant drawbacks: computational inefficiency and lack of interpretability into important features for differentiating clusters. We address these two challenges by developing IMPACC: Interpretable MiniPatch Adaptive Consensus Clustering. We ensemble cluster co-occurrences from tiny subsets of both observations and features, termed minipatches, thus dramatically reducing computation time. Additionally, we develop adaptive sampling schemes for observations, which result in both improved reliability and computational savings, as well as adaptive sampling of features, which leads to interpretable solutions by quickly learning the most relevant features that differentiate clusters. We study our approach on synthetic data and a variety of real large-scale bioinformatics data sets; results show that our approach not only yields more accurate and interpretable cluster solutions, but also substantially improves computational efficiency compared to standard consensus clustering approaches.

E0710: **Universal rank inference via residual subsampling with application to large networks**
Presenter: **Yingying Fan**, University of Southern California, United States
Co-authors: Xiao Han, Qing Yang

Determining the precise rank is an important problem in many large-scale applications with matrix data exploiting low-rank plus noise models. We suggest a universal approach to rank inference via residual subsampling (RIRS) for testing and estimating rank in a wide family of models, including many popularly used network models such as the degree corrected mixed membership model as a special case. The procedure constructs a test statistic via subsampling entries of the residual matrix after extracting the spiked components. The test statistic converges in distribution to the standard normal under the null hypothesis, and diverges to infinity with asymptotic probability one under the alternative hypothesis. The effectiveness of RIRS procedure is justified theoretically, utilizing asymptotic expansions of eigenvectors and eigenvalues for large random matrices recently developed. The advantages of the newly suggested procedure are demonstrated through several simulations and real data examples.

E1620: **A sparse random graph model for sparse directed networks**
Presenter: **Stefan Stein**, University of Warwick, United Kingdom
Co-authors: Chenlei Leng

An increasingly urgent task in the analysis of networks is to develop statistical models that include contextual information in the form of covariates while respecting degree heterogeneity and sparsity. We propose a new parameter-sparse random graph model for density-sparse directed networks, with parameters to explicitly account for all these features. The resulting objective function of our model is akin to that of the high-dimensional logistic regression, with the key difference that the probabilities are allowed to go to zero at a certain rate to accommodate sparse networks. We show that under appropriate conditions, an estimator obtained by the familiar penalized likelihood with an ℓ_1 penalty to achieve parameter sparsity can alleviate the curse of dimensionality, and crucially is selection and rate consistent. Interestingly, inference on the covariate parameter can be conducted straightforwardly after the model fitting, without the need of the kind of debiasing commonly employed in ℓ_1 penalized likelihood estimation. In developing our model, we provide the first result highlighting the fallacy of what we call data-selective inference, a common practice of artificially truncating the sample by throwing away nodes based on their connections, by examining the estimation bias in the Erdös-Rényi model theoretically and in the stochastic block model empirically.

E1051: **The multirank likelihood for semiparametric CCA**
Presenter: **Jordan Bryan**, Duke University, United States
Co-authors: Peter Hoff

In the analysis of multivariate data, it is often of interest to evaluate the dependence between two or more sets of variables rather than the dependence between individual variables. Canonical correlation analysis (CCA) is a classical data analysis technique that estimates parameters describing the dependence between such sets. However, inference procedures based on traditional CCA rely on the assumption that all variables are jointly normally distributed. We present a semiparametric approach to CCA in which the multivariate margins of each variable set may be arbitrary, whereas the dependence between variable sets is still described by a parametric model that provides a low dimensional summary of dependence. This approach is a generalization of the Gaussian copula model to the case of vector-valued margins. While maximum likelihood estimation in the proposed model is intractable, we develop a novel MCMC algorithm called cyclically monotone Monte Carlo (CMMC) that allows us to get estimates and confidence regions for the between-set dependence parameters. This algorithm is based on a multirank likelihood function, which uses only part of the information in the observed data in exchange for being free of assumptions about the multivariate margins. We illustrate the proposed inference procedure on nutrient data from the USDA.

EO746 Room Virtual R35 COMPUTATIONAL STATISTICAL METHODS FOR ENVIRONMENTAL SCIENCES Chair: Seiyon Lee

E1152: **Fast expectation-maximization algorithms for spatial generalized linear mixed models**
Presenter: **Yawen Guan**, University of Nebraska - Lincoln, United States
Co-authors: Murali Haran

Spatial generalized linear mixed models (SGLMMs) are popular and flexible models for non-Gaussian spatial data. They are useful for spatial interpolations as well as for fitting regression models that account for spatial dependence, and are commonly used in many disciplines such as epidemiology, atmospheric science, and sociology. Inference for SGLMMs is typically carried out under the Bayesian framework at least in part because computational issues make maximum likelihood estimation challenging, especially when high-dimensional spatial data are involved. We provide a computationally efficient projection-based maximum likelihood approach and two computationally efficient algorithms for routinely fitting SGLMMs. The two algorithms proposed are both variants of the expectation-maximization algorithm, using either Markov chain Monte Carlo or a Laplace approximation for the conditional expectation. The methodology is general and applies to both discrete-domain (Gaussian Markov random field) as well as continuous-domain (Gaussian process) spatial models. We show, via simulation and real data applications, that our methods perform well both in terms of parameter estimation as well as prediction. Crucially, our methodology is computationally efficient and scales well with the size of the data and is applicable to problems where maximum likelihood estimation was previously infeasible.

E1258: **Spatial scale-aware tail dependence modeling for high-dimensional spatial extremes**
Presenter: **Likun Zhang**, Lawrence Berkeley National Lab, United States
Co-authors: Mark Risser, Benjamin Shaby

Extreme events over large spatial domains like the contiguous United States may exhibit highly heterogeneous tail dependence characteristics, yet most existing spatial extremes models yield only one dependence class over the entire spatial domain. To accurately characterize storm dependence in analysis of extreme events, we propose a mixture component model that achieves flexible dependence properties and allows truly high-dimensional inference for extremes of spatial processes. We modify the popular random scale construction that multiplies a Gaussian random field by a single radial variable; that is, we add non-stationarity to the Gaussian process while allowing the radial variable to vary smoothly across space. As the level of extremeness increases, this single model exhibits both long-range asymptotic independence and short-range weakening

dependence strength that leads to either asymptotic dependence or independence. Under the assumption of local stationarity, we make inferences on the model parameters using local Bayesian hierarchical models, and run adaptive Metropolis algorithms concurrently via parallelization. Then, after conducting posterior inference locally, the mixture component representation of the model coherently ties the local posteriors together to obtain a globally nonstationary model.

E1270: **Flexible basis representations for modeling high-dimensional hierarchical spatial data using adaptive resolution tuning**
Presenter: **Remy MacDonald**, George Mason University, United States
Co-authors: Seiyon Lee

Non-Gaussian spatial data are prevalent across many fields (e.g., animal counts in ecology, count data on disease incidence, pollutant concentrations near highways, and the incidence of cloud cover in satellite imagery). Spatial generalized linear mixed models are a highly flexible class of spatial models for non-Gaussian spatial data, but these can be computationally prohibitive for large datasets. To address this challenge, past studies approximate the spatial random field using basis functions; thereby exploiting the low-rank structures and bypassing large matrix operations. Popular basis representation methods employ nested radial basis functions with fixed knot locations and bandwidths, but these must be fixed a priori and this specification affects model performance. We propose a data-driven, adaptive algorithm that results in increased flexibility and fast model-fitting: (1) the knot locations are selected based on a space-covering design; (2) we partition the spatial domain into disjoint subregions such that the smoothing parameter varies across partitions; and (3) in our Bayesian model, the smoothing parameters are allowed to vary. Our approach extends to a wide array of spatial data (e.g. non-stationary, stationary, or non-Gaussian) and results in strong predictive ability. We demonstrate our method through simulation studies and applications to real-world datasets.

E1291: **Continuous-time discrete-space (CTDS) movement models over two- and three-dimensional space**
Presenter: **Joshua Hewitt**, Duke University, United States
Co-authors: Robert Schick, Alan Gelfand

Continuous-time discrete-space (CTDS) processes for animal movement model trajectories across discretely observed spatial domains, which arise via gridded remote sensing products in 2D, or via underwater sound propagation models in 3D. CTDS models are approximately estimated from finite observations of animal location because the exact likelihood has $O(N^3)$ computational complexity, where N is the size of the spatial domain. The usual approximation averages estimates from multiple imputations of the complete, unobserved trajectory. However, imputations typically discretize output from continuous-space surrogate models that do not account for complex boundaries like coastlines and bathymetry. As a result, parameter and location estimates may be biased by imputations that are inconsistent with respect to physical barriers. We remedy this issue by proposing a discrete-time likelihood approximation. We also develop additional theory and interpretation for CTDS model formulation, which supports the approximation. We demonstrate the improvement of the discrete-time approximation in simulation. We also demonstrate improvement in an application, by combining GPS locations and depth measurements with bathymetry data to refine estimates of a beaked whales position during exposure to anthropogenic sound.

EO748 Room Virtual R36 STATISTICAL THEORY FOR MACHINE LEARNING METHODS Chair: Michael Vogt

E0384: **Evading the curse of dimensionality in nonparametric regression with deep neural networks**
Presenter: **Sophie Langer**, University Twente, Germany

In the classical multivariate regression context, it is well-known that any nonparametric method is affected by the so-called curse of dimensionality, meaning that convergence of the estimators slows down as dimension increases. We show how one can evade this phenomenon by using estimators based on deep neural networks and restricting the class of regression functions in a proper sense. In a second result, we consider the case that the predictor variable is concentrated on a manifold and show again that deep neural network estimators achieve a convergence rate independent of dimension.

E0887: **Benign overfitting in distributed learning**
Presenter: **Nicole Muecke**, TU Braunschweig, Germany

While large training datasets generally offer improvement in model performance, the training process becomes computationally expensive and time-consuming. Distributed learning (DL) is a common strategy to reduce the overall training time by exploiting multiple computing devices. Recently it has been observed in the single machine setting that overparameterization is essential for benign overfitting. We analyze distributed ridgeless linear regression and show that overparameterization is essential for benign overfitting also in the distributed setting. Moreover, we show that both the covariance structure and the hardness of the learning expressed in terms of a source condition determine the efficiency of DL in the presence of overparameterization. The results show that the number of local nodes acts as an explicit regularization parameter and the efficiency may increase linearly with the number of machines. This is in stark contrast to the underparameterized regime where the efficiency is known to be decreasing linearly.

E0970: **Graphical models for stationary time series**
Presenter: **Sumanta Basu**, Cornell University, United States

Graphical models offer a powerful framework to capture intertemporal and contemporaneous relationships among the components of a stationary multivariate time series. These relationships are encoded in the multivariate spectral density matrix and its inverse. We will present adaptive thresholding and penalization methods for the estimation of these objects under suitable sparsity assumptions. We will discuss new optimization algorithms and investigate the consistency of estimation under a double-asymptotic regime where the dimension of the time series increases with sample size.

E1277: **Average derivative estimation with bayesian decision tree ensembles**
Presenter: **Christoph Breunig**, Emory University, United States

The estimation of average derivatives with Bayesian decision tree ensembles is considered. We make use of soft decision trees in which the decisions are treated as probabilistic. Specifically, we make use of the Bayesian additive regression trees framework. The posterior distribution concentrates at the minimax rate for estimating directional derivatives (up to a logarithmic factor) for sparse functions and functions with additive structures in the high dimensional regime. We illustrate the finite sample properties in simulations and empirical illustrations.

EO533 Room Virtual R37 PROJECTION PURSUIT II Chair: Nicola Loperfido

E0861: **Portfolio optimization by a bivariate functional of the mean and variance**
Presenter: **Zinoviy Landsman**, University of Haifa, Israel

The focus is on the problem of maximization of a functional of the expected portfolio return and variance portfolio return in its most general form and the presentation of an explicit closed-form solution of the optimal portfolio selection. This problem is closely related to the expected utility maximization and the two-moment decision models. We show that the most-known risk measures, such as mean-variance, expected short-fall, Sharpe ratio, generalized Sharpe ratio and the recently introduced tail mean-variance, are special cases of this functional. The new results essentially generalize previous results concerning the maximization of a combination of the expected portfolio return and a function of the variance of portfolio return. The general mean-variance functional is not restricted to a concave function with a single optimal solution. Thus, we also provide optimal solutions to a fractional programming problem, that is arising in portfolio theory. The obtained analytic solution of the optimization problem allows

us to conclude that all the optimization problems corresponding to the general functional have efficient frontiers belonging to the efficient frontier obtained for the mean-variance portfolio.

E0889: **The tail risk projection: From multivariate risk measures to projection pursuit**
Presenter: **Tomer Shushi**, Ben Gurion University of the Negev, Israel
The Tail Risk Projection (TRP) is defined and applications are shown for producing multivariate risk measures that take into account the dependence structure of the risks when the focus on extreme loss events. The multivariate risk measures could facilitate a systemic risk measure with explicit expressions for exponential dispersion models subject to any pre-specified systemic event. We then examine the use of Projection Pursuit as part of the risk measurement process.

E0567: **Section pursuit**
Presenter: **Ursula Laa**, BOKU University, Austria
Co-authors: Di Cook, Andreas Buja, German Valencia
Multivariate data is often visualized using linear projections, produced by techniques such as principal component analysis, linear discriminant analysis, and projection pursuit. A problem with projections is that they obscure low and high-density regions near the center of the distribution. Sections, or slices, can help to reveal them. Section pursuit (building on the extensive work in projection pursuit) is introduced, a new method to search for interesting slices of the data. Linear projections are used to define sections of the parameter space, and we calculate interestingness by comparing the distribution of observations, inside and outside a section. By optimizing this index, it is possible to reveal features such as holes (low density) or grains (high density), which can be useful when data distributions depart from uniform or normal, as in visually exploring nonlinear manifolds, and functions in multivariate space. We will show how section pursuit can be applied when exploring decision boundaries from classification models or when exploring subspaces induced by complex inequality conditions from a multiple parameter model.

E0835: **Asymmetric projection pursuit for classification**
Presenter: **Nicola Loperfido**, University of Urbino, Italy
When applied to cluster analysis, projection pursuit aims at finding the data projections which best separate clusters. Asymmetric projection pursuit for classification finds the direction which best separates one cluster from the remaining data, removes the cluster and repeats the previous steps until no clusters are left. The chosen projection pursuit index is kurtosis, which is particularly apt at data dichotomization. Asymmetric projection pursuit builds upon asymmetric linear dimension reduction for classification. The method is theoretically motivated using finite normal mixtures and it is empirically illustrated with the Iris dataset.

EO452 Room K2.31 Nash (Hybrid 07) SPATIO-TEMPORAL MODELING OF INFECTIOUS DISEASES (VIRTUAL) Chair: Andrew Lawson

E0470: **Behavioural change in spatial epidemic models**
Presenter: **Rob Deardon**, University of Calgary, Canada
One of the numerous difficulties in modelling epidemic spread is that caused by a behavioural change in the underlying population. This can be a major issue in public health since, as we have seen during the COVID-19 pandemic, behaviour in the population can change drastically as infection levels vary, both due to government mandates and personal decisions. Such changes in the underlying population result in major changes in transmission dynamics of the disease, making the modelling challenges. However, these issues arise in agriculture and public health, as changes in farming practice are also often observed as disease prevalence changes. We consider whether it is possible to model the behavioural change mechanism and the underlying transmission dynamics in the context of spatial epidemic models. Multiple mechanisms for allowing behavioural change will be considered, and issues such as identifiability for such models be considered. Models will be fitted within a Bayesian framework and explored using simulated data.

E0882: **Comparison of Bayesian spatio-temporal models in the infectious disease outbreak surveillance**
Presenter: **Joanne Kim**, Medical University of South Carolina, United States
Co-authors: Andrew Lawson
COVID-19 pandemic raises the awareness of the public health community for the surveillance of infectious disease outbreaks. Bayesian spatio-temporal models for small area health data have been widely used to analyze infectious disease progression. These models were used for retrospective analysis and prospective surveillance analysis. Retrospective analysis is commonly used to describe the infectious disease outbreak phenomenon; on the other hand, prospective surveillance analysis is more focused on the detection of abnormal patterns of the current data. We compare spatio-temporal models in both retrospective and prospective analysis settings and present and discuss the result for their goodness of fit and prediction ability. The analysis result for both simulation and COVID-19 incidence data of several US states are presented.

E1141: **A latent spatial model for pandemic prediction**
Presenter: **Dani Gamerman**, Universidade Federal do Rio de Janeiro, Brazil
Co-authors: Marcos Prates, Samuel Faria, Mauricio Castro
Epidemic modeling consists of the specification of an underlying structure that could rely entirely on epidemiological reasoning, be data-driven or a combination of them. In any case, it is based on the identification of characteristics that are shared by many regions. Some of these features present similarities across observational units. Hierarchical modeling is particularly useful in these settings as it allows the explicit incorporation of these similarities, thus enabling borrowing information across regions. The resulting setup is suitable for the estimation of the epidemic evolution and prediction of future epidemic cases. A number of options are considered, including those taking spatial configuration into account. These ideas are illustrated in the analysis of the evolution of Covid19 in Brazil, integrated across its 27 states.

E1406: **Global space-time mapping of antimicrobial-resistance among selected priority bacterial pathogens, 2000-2019**
Presenter: **Benn Sartorius**, University of Oxford, United Kingdom
Co-authors: Annie Browne, Michael Chipeta, Frederick Fell, Sean Hackett, Georgina Haines-Woodhouse, Bahar Kashef Hamadani, Emmanuelle Kumaran, Catrin Moore, Christiane Dolecek
Antimicrobial resistance (AMR) is a major and growing public health threat. Despite considerable global efforts, our understanding of AMR burden across space-time remains sparse. Understanding this is essential to better inform policy and help combat the further spread of AMR. The flagship Global Research on AntiMicrobial resistance (GRAM) Project, attempting to improve our understanding of global AMR burden, has compiled proportions of resistant isolates for priority bacterial pathogens to key antimicrobials by country/year. These data currently span 2500 sources including patient-level AMR microbiology data, aggregated AMR data (e.g. surveillance), published studies and directly from collaborators. Influential covariates for each bacterial-antimicrobial combination were included in a machine learning stacked ensemble framework to improve predictive validity. We adapted and employed Bayesian space-time Gaussian Process Regression and conditional autoregressive modelling to help improve estimates in countries without data and uncertainty interval quantification. Initial results show large variations in resistance by pathogen, antimicrobial, location and year. Notably, particular bacteria-antimicrobial combinations have higher and/or increasing burden in many LMIC over the period. This will have important policy implications, especially in the context of the current COVID pandemic. The next stages will include refining/optimising the modelling framework.

EO398 Room K2.40 (Hybrid 08) RECENT ADVANCES IN BAYESIAN APPROACHES TO NEUROIMAGING Chair: Aaron Scheffler

E0473: **A Bayesian regression framework for brain imaging data with multiple structural- and network-valued predictors**
Presenter: **Aaron Scheffler**, University of California, San Francisco, United States
Co-authors: Rajarshi Guhaniyogi, Rene Marquez

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. Still, this integration can be hindered by the data complexity. Each image contains structural information, indexing spatial information, or network information, indexing connectivity among the image, which reinforce each other but are challenging to merge. We propose a Bayesian regression framework that provides inference and prediction for a scalar outcome as a function of a multi-object predictor. Our framework will accommodate multiple image predictors with different structures and identify image regions that influence the response jointly via efficient hierarchical prior structures that scale to high-resolution image data volume. A working example is provided to predict language comprehension scores from multi-object image data to explore the neural underpinnings of language loss in primary progressive aphasia patients.

E0613: **Bayesian image analysis in Fourier space models and some relationships with Markov random fields**
Presenter: **John Kornak**, University of California, San Francisco, United States
Co-authors: Karl Young, Eric Friedman

Bayesian image analysis can improve image quality by balancing a priori expectations of image characteristics with a model for the noise process. The conventional Bayesian model in the image space approach implements priors that describe inter-dependence between spatial locations. It can therefore be difficult to model and compute. However, similar models can be developed more conveniently as a large set of independent processes when considered in Fourier space. The originally complex high-dimensional estimation problem in image space is thereby broken down into a series of (trivially parallelizable) independent one-dimensional problems in Fourier space. Example implementations of a range Bayesian image analysis in Fourier space models will be shown. How these Fourier space models relate to Markov random field-based models that are commonly used in conventional Bayesian image analysis will also be discussed.

E0730: **New ideas on Bayesian data sketching**
Presenter: **Rajarshi Guhaniyogi**, Texas A & M university, United States
Co-authors: Aaron Scheffler

Bayesian computation of high dimensional linear regression models with a popular Gaussian scale mixture prior distribution using Markov Chain Monte Carlo (MCMC) or its variants can be extremely slow or completely prohibitive due to the heavy computational cost that grows in the cubic order with the number of features. We adopt the data sketching approach to compress the original samples by a random linear transformation to m samples, and compute Bayesian regression with Gaussian scale mixture prior distributions with the randomly compressed response vector and feature matrix. The proposed approach yields computational complexity growing in the cubic order of m. Another important motivation for this compression procedure is that it anonymizes the data by revealing little information about the original data in the course of analysis. The detailed empirical investigation with the Horseshoe prior from the class of Gaussian scale mixture priors shows closely similar inference and a massive reduction in per iteration computation time of the proposed approach compared to the regression with the full sample. We characterize the dimension of the compressed response vector m as a function of the sample size, the number of predictors and the sparsity in the regression to guarantee accurate estimation of predictor coefficients asymptotically, even after data compression.

E1220: **Bayesian time-varying tensor vector autoregressive models for dynamic effective connectivity**
Presenter: **Michele Guindani**, University of California, Irvine, United States

Recent developments in neuroimaging investigate how some brain regions directly influence the activity of other regions of the brain dynamically throughout the course of an experiment. Time-varying vector autoregressive (TV-VAR) models have been employed to draw inferences for this purpose, but they are very computationally intensive since the number of parameters to be estimated increases quadratically with the number of time series. We propose a computationally efficient Bayesian time-varying VAR approach for modeling high-dimensional time series. The proposed framework employs a tensor decomposition for the VAR coefficient matrices at different lags. Dynamically varying connectivity patterns are captured by assuming that at any given time only a subset of components in the tensor decomposition is active. Latent binary time series select the active components at each time via a convenient Ising prior specification. The proposed prior structure encourages sparsity in the tensor structure and allows to ascertain model complexity through the posterior distribution. More specifically, sparsity-inducing priors are employed to allow for global-local shrinkage of the coefficients, to determine automatically the rank of the tensor decomposition, and to guide the selection of the lags of the auto-regression. We show the performances of our model formulation via simulation studies and data from a real fMRI study involving a book reading experiment.

EG021 Room Virtual R18 CONTRIBUTIONS IN BAYESIAN STATISTICS I Chair: Antonio Canale

E0953: **Baseline methods for estimating the parameters of Generalized Pareto distributions**
Presenter: **Eva Lopez Sanjuan**, Universidad de Extremadura, Spain
Co-authors: M Isabel Parra Arevalo, Mario Martinez Pizarro, Jacinto Martin Jimenez

In the parameter estimation of limit extreme value distributions, most employed methods only use some of the available data. Using the peaks-over-threshold method for Generalized Pareto distribution (GPD), only the observations above a certain threshold are considered. To make the most of the information provided by the observations, and improve the accuracy in Bayesian parameter estimation, we present two new Bayesian methods to estimate the parameters of the GPD. They take into account the whole data set from the baseline distribution, and the existing relations between the baseline and the limit GPD parameters, in order to define highly informative priors. We make a comparison between the Bayesian Metropolis-Hastings algorithm with data over the threshold and the new methods when baseline distribution is a stable distribution, whose properties assure we can reduce the problem to study standard distributions and also allow us to propose new estimators for the parameters of the tail distribution. Specifically, three cases of stable distributions were considered: Normal, Levy and Cauchy distributions. Nevertheless, the methods would be applicable to many other baseline distributions, through finding relations between baseline and GPD parameters via studies of simulations.

E1311: **A Bayesian approach for combining probability and non-probability samples for analytic inference**
Presenter: **Camilla Salvatore**, University of Milano-Bicocca, Italy
Co-authors: Silvia Biffignandi, Joseph Sakshaug, Arkadiusz Wisniowski, Bella Struminskaya

The popularity of non-probability sample web-surveys is increasing due to their convenience and relatively low costs. On the contrary, traditional probability-sample surveys are suffering from decreasing response rates, with a consequent increase in survey costs. Integrating the two samples in order to overcome their respective disadvantages is one of the current challenges in the statistical field. Our aim is to combine probability and non-probability samples to improve analytic inference on model parameters. We consider the Bayesian framework, where inference is based on a probability sample and available information about a non-probability sample is provided naturally through the prior. We focus on the logistic regression case and conduct a simulation study under different scenarios based on selection variables, selection probabilities, sample sizes, and prior specifications. We compare the performance of several informative and non-informative priors in terms of mean-squared errors (MSE). Overall,

the informative priors reduce the MSE or, in the worst-case situation perform equivalently to the non-informative priors. Finally, we present a real data analysis considering an actual probability-based survey and several volunteer web-surveys which represent different selection scenarios.

E1669: **The taxicab sampler: MCMC for discrete spaces with application to tree models**
Presenter: **Vincent Geels**, The Ohio State University, United States
Co-authors: Matthew Pratola, Radu Herbei
Motivated by the problem of exploring discrete but very complex state spaces in Bayesian models, a novel Markov Chain Monte Carlo search algorithm is proposed: the taxicab sampler. We describe the construction of this sampler and discuss how its interpretation and usage differs from that of standard Metropolis-Hastings as well as the closely-related Hamming ball sampler. The proposed taxicab sampling algorithm is then shown to demonstrate substantial improvement in computation time relative to a naive Metropolis-Hastings search in a motivating Bayesian regression tree count model, in which we leverage the discrete state space assumption to construct a novel likelihood function that allows for flexibly describing different mean-variance relationships while preserving parameter interpretability compared to existing likelihood functions for count data.

E1643: **A Bayesian multilevel hidden Markov model with time-varying covariates for multivariate count time series**
Presenter: **Sebastian Mildiner Moraga**, Utrecht University, Netherlands
Co-authors: Emmeke Aarts
Technological advances such as accelerometers, tracking devices, automatic coding of video recordings, and in vivo experimental set-ups made it easier and more affordable to collect data on multiple subjects or animals with a high resolution over time. However, the high dimensionality of these data combined with their nested structure makes them challenging to analyse. Moreover, extracting insights about the dynamics of processes underlying these data also requires statistical models capable of tracking changes on multiple individuals across time. The hidden Markov models (HMMs) are a promising approach to this end, as they can summarize complex processes with a set of hidden states that switch over time. We present a novel parametric multilevel HMM (MHMM) with continuously distributed random effects and a Poisson-LogNormal emission distribution. Our model deals with the nested structure of the data and allows for including time-varying covariates in the transition distribution. We illustrate the use of our model with a small simulation based on an empirical dataset with multi-electrode electrophysiological measurements in monkeys. In addition, we show how the MHMM can be used to obtain the forward probabilities of the states to investigate changes in a latent process over time.

CO539 Room Virtual R33 EMPIRICAL ASPECTS OF CRYPTOCURRENCY MARKETS Chair: Pierangelo De Pace

C0712: **Measuring cryptocurrency price co-movement using a thick pen**
Presenter: **Seungho Lee**, University of Aberdeen, United Kingdom
Co-authors: Marc Gronwald, Sania Wadud, Robert B Durand, Yuan Zhao
Integration of the cryptocurrency markets is measured by using two so-called Thick Pen methods: the Thick Pen Measure of Association (TPMA) as well as Multi-Thickness Thick Pen Measure of Association (MTTPMA). They allow one to capture time-varying co-movement of different cryptocurrency prices as well as co-movement at different time scales; i.e. short-term and long-term features of the price series. A particular strength in this application is the ability to detect instabilities in price relationships. The analysis shows that there is strong co-movement between the price series under consideration. These findings are also critically discussed as cryptocurrencies are characterised by different monetary policies and, thus, are not identical entities.

C0987: **Heterogeneous agents and cryptocurrency**
Presenter: **Marco Lorusso**, Newcastle University, United Kingdom
Co-authors: Francesco Ravazzolo, Stefano Grassi
A Dynamic Stochastic General Equilibrium (DSGE) model with heterogeneous agents is developed and estimated in which a first group of agents participates in the cryptocurrency market, whereas a second group of agents does not hold a cryptocurrency. We estimate our model using a Mixture of Students t by Importance Sampling weighted Expectation-Maximization (MitISEM). The use of Importance Sampling and of an adaptive scheme based on Expectation-Maximization allows us to efficiently estimate our heterogeneous agent macro model with aggregate shocks. Results indicate that heterogeneity matters for the magnitude of crypto-specific shocks to the economy.

C1090: **Investors' beliefs and cryptocurrency prices**
Presenter: **Matteo Benetton**, Berkeley Haas, United States
Co-authors: Giovanni Compiani
The aim is to explore the impact of investors' beliefs on cryptocurrency demand and prices using three individual-level surveys and a structural characteristics-based demand model with differentiated cryptocurrencies and heterogeneous investors. We show that younger individuals with lower income are more optimistic about the future value of cryptocurrencies, as are late investors. We identify the model combining observable beliefs with an instrumental variable strategy that exploits variation in the production of different cryptocurrencies. Counterfactual analyses quantify the impact on equilibrium prices and portfolio allocations of (i) entry of late optimistic investors, and (ii) growing concerns about the sustainability of energy-intensive proof-of-work cryptocurrencies.

C1457: **Diversification among cryptoassets: Bitcoin maximalism, active portfolio management, and survival bias**
Presenter: **Ladislav Kristoufek**, Institute of Information Theory and Automation, Czech Academy of Sciences, Czech Republic
Cryptoassets and mostly Bitcoin have attracted the attention of institutional investors in the latest price rallies of 2020 and 2021. The need for other cryptoassets apart from Bitcoin in their portfolios is mostly unexplored in the current literature and the general perception of diversification benefits within crypto-markets mostly builds on popular beliefs. The current study delivers a deep dive into active and passive investment strategies, focusing on specifics of cryptoassets, most importantly the survival bias in the portfolio dataset construction and its implications. We show that the survival bias is in fact driving the results at their very core and the differences between using the backwards-looking subset of assets and the actual assets available at the time of portfolio construction are substantial and lead to completely different implications and investment suggestions. It turns out that active portfolio management does not pay off in most instances compared to simply holding Bitcoin.

CO164 Room Virtual R34 IMPULSE RESPONSES Chair: Michael Owyang

C0238: **Local projections, autocorrelation, and efficiency**
Presenter: **Amaze Lusompa**, Federal Reserve Bank of Kansas City, United States
It is well known that Local Projections (LP) residuals are autocorrelated. Conventional wisdom says that LP have to be estimated by OLS with Newey-West (or some type of Heteroskedastic and Autocorrelation Consistent (HAC)) standard errors and that GLS is not possible because the autocorrelation process is unknown and/or because the GLS estimator would be inconsistent. We show that the autocorrelation process of LP is known and can be corrected by using a consistent GLS estimator. Estimating LP with GLS has three major implications: 1) LP GLS can be less biased, more efficient, and generally has better coverage properties than estimation by OLS with HAC standard errors. 2) Consistency of the LP GLS estimator gives a general counterexample showing that strict exogeneity is not a necessary condition for GLS. 3) Since the autocorrelation process can be modeled explicitly, it is now possible to estimate time-varying parameter LP.

C0405: **Identifying high-frequency shocks in nonlinear models**
Presenter: **Alessia Paccagnini**, University College Dublin, Ireland
Co-authors: Fabio Parla

Focusing on mixed frequency regressions, a novel approach implemented in a nonlinear setting is proposed. Applying Bayesian techniques, we identify high-frequency shocks studying the behavior of the temporal aggregation bias across the business cycle phases. Montecarlo experiments provide an assessment of the results using different Data Generation Processes. The empirical illustration provides new evidence to identify the effects of the uncertainty shock on the US economy's macro variables disentangling between normal times, recessionary times, and disaster events.

C1059: **Impulse response analysis for structural dynamic models with nonlinear regressors**
Presenter: **Ana Maria Herrera**, University of Kentucky, United States
Co-authors: Elena Pesavento, Silvia Goncalves, Lutz Kilian

The construction of nonlinear impulse responses in linear structural dynamic models that include nonlinearly transformed regressors is studied. We derive the closed-form solution for the population impulse responses to a given shock and propose a control function approach to estimating these responses without taking a stand on how the remainder of the model is identified. Our plug-estimator dispenses with the need for simulations and, unlike conventional local projection (LP) estimators, is consistent. A modified LP estimator is shown to be consistent in special cases, but less accurate in finite samples than the plug-in estimator.

C0494: **Impulse response functions in a self-exciting world**
Presenter: **Daniel Soques**, University of North Carolina Wilmington, United States
Co-authors: Neville Francis, Michael Owyang

Impulse response functions are calculated in situations where the data generating process depends upon a regime process that is driven by model variables. Researchers commonly use either generalized impulse response functions or local projections to estimate the true impulse response in such self-exciting models. A tradeoff exists between these two methods: local projections depend on the observed switches in the data while generalized impulse responses rely on specifying the correct regime process. We investigate under which conditions (i.e., model misspecification) each method is preferred. We then revisit the estimation of the fiscal multiplier using each method.

CO040 Room Virtual R38 ECONOMETRIC METHODS FOR HIGH-FREQUENCY DATA Chair: Massimiliano Caporin

C0237: **A modern take on market efficiency: The impact of Trump's tweets on financial markets**
Presenter: **Zorka Simon**, Leibniz Institute for Financial Research SAFE, Germany
Co-authors: Farshid Abdi, Emily Kormanyos, Loriana Pelizzon, Mila Getmansky Sherman

The focus is on the role of social media as a high-frequency, unfiltered mass information transmission channel and how its use for government communication affects the aggregate stock markets. To measure this effect, we concentrate on one of the most prominent Twitter users, the 45th President of the United States, Donald J. Trump. We analyze around 1,400 of his tweets related to the US economy and classify them by topic and textual sentiment using machine learning algorithms. We investigate whether the tweets contain relevant information for financial markets, i.e. whether they affect market returns, volatility, and trading volumes. Using high-frequency data, we find that Trump's tweets are most often a reaction to pre-existing market trends and therefore do not provide material new information that would influence prices or trading. We show that past market information can help predict Trump's decision to tweet about the economy.

C0241: **Not all words are equal: Sentiment and jumps in cryptocurrency market**
Presenter: **Francesco Poli**, University of Padova, Italy
Co-authors: Massimiliano Caporin, Oguzhan Cepni, Ahmet Faruk Aysan

The aim is to decipher the relation between price jumps and investors' sentiment on cryptocurrencies. We use intraday (1-minute) data on a large set of cryptocurrencies prices and Thomson Reuters MarketPysch Indices (TRMIs) which are monitoring sentiment on the same cryptocurrencies by applying complex natural language processing to news and social media. We detect jumps at the intra-day level and correlate their occurrence with TRMI-scored events through logistic regressions. Our results show moderate evidence that the release of information, as monitored by the TRMIs, increases the probability of both positive and negative price jumps, especially within 60 min from the TRMI event. In particular, we find that while sentiment on topics limited to emotions such as optimism, anger, conflict and fear increases the probability of positive jumps occurrence, a more extensive set of topics including both emotional factors and risks perceptions related to volatility, scam, violence and price forecast is identified as the possible causation of negative jumps. In addition, the size of the jumps seems to be less related to the TRMIs compared to the jump occurrence, which has some impact on the occurrence of further TRMIs events. Furthermore, our findings suggest the presence of a significant self-excitation, that is, the occurrence of price jumps (TRMIs events) strongly increases the probability of the occurrence of price jumps (TRMI events).

C0504: **Testing for endogeneity of irregular sampling schemes**
Presenter: **Davide Pirino**, University of Rome Tor Vergata, Italy
Co-authors: Aleksey Kolokolov, Giulia Livieri

In the context of high-frequency data, a simplifying assumption of the independence between the sampling scheme and the observed process itself is often made. In order to justify or reject the assumption empirically, we propose a statistical test for the endogeneity of the sampling times. The test is robust to the presence of jumps and can be used for detecting the dependence between zeros in the financial prices sampled at a moderate frequency and the efficient price process. Extensive Monte Carlo simulations confirm the good finite sample performance of the proposed test.

C0738: **A model-based spillover index**
Presenter: **Massimiliano Caporin**, University of Padova, Italy
Co-authors: Giuseppe Storti

A generalized version of the spillover index is derived from a conditional autoregressive Wishart model (WAR). The WAR model is first estimated by Quasi Maximum Likelihood under L1-penalization and taking advantage of the analytical gradient. The spillover index is then computed, accounting for the interdependence between realized variances and covariances. To recover the index, a novel rectangular forecast error variance decomposition is introduced, assuming shocks on N stock returns and a reaction on the N realized variances and the $0.5N(N-1)$ realized covariances. Empirical examples contrast our index, with and without interdependencies in the realized covariance modeling, to the original parameterization based on a Vector Auto-Regressive (VAR) model including only realized variances.

CO362 Room Virtual R39 TIME SERIES ECONOMETRICS Chair: Josu Arteche

C0572: **Single step estimation of ARMA roots for non-fundamental nonstationary fractional models**
Presenter: **Carlos Velasco**, Universidad Carlos III de Madrid, Spain
Co-authors: Ignacio Lobato

A single step estimator is proposed for the autoregressive and moving-average roots(without imposing causality or invertibility restrictions) of a nonstationary Fractional ARMA process. These estimators employ an efficient tapering procedure, which allows for a long memory component in the process, but avoid estimating the nonstationarity component, which can be stochastic and/or deterministic. After selecting automatically the order of the model, we robustly estimate the AR and MA roots for trading volume for the thirty stocks in the Dow Jones Industrial Average Index

in the last decade. Two empirical results are found. First, there is strong evidence that stock market trading volume exhibits non-fundamentalness. Second, non-causality is more common than non-invertibility.

C0758: **Local bootstrap in long memory time series**
Presenter: **Josu Arteche**, University of the Basque Country UPV/EHU, Spain

Bootstrap techniques in the frequency domain have been proved to be effective instruments to approximate the distribution of many statistics of weakly dependent (short memory) series, although their validity with long memory remains unsolved. A Frequency Domain Local Bootstrap (FDLB) is proposed based on resampling a locally studentised version of the periodogram in a neighbourhood of the frequency of interest. We analyse the similarities of the distribution of the periodogram and the FDLB distribution in stationary and non-stationary long memory series. A bound of the Mallows distance between the distributions of the original and bootstrap periodograms is offered for stationary and non-stationary long memory series. This result is in turn used to justify the use of FDLB for some statistics such as the average periodogram or the Local Whittle (LW) estimator. Finally, the finite sample behaviour of the FDLB in the LW estimator is analysed in a Monte Carlo, comparing its performance with rival alternatives.

C0963: **Testing for multiple structural breaks in multivariate long memory time series**
Presenter: **Philipp Sibbertsen**, University of Hannover, Germany
Co-authors: Vivien Less

Estimation and testing of multiple breaks that occur at unknown dates are considered in multivariate long-memory time series. We propose a likelihood ratio based approach for estimating breaks in the mean and the covariance of a system of long-memory time series. The limiting distribution of these estimates as well as the consistency of the estimators are derived. A testing procedure to determine the unknown number of break points is given based on iterative testing on the regression residuals. A Monte Carlo exercise shows the finite sample performance of our method. An empirical application to inflation series illustrates the usefulness of our procedures.

C1324: **A state-space frame for modelling time-varying parameters in panels: An applicacion to the Okun's law**
Presenter: **Juan Sapena**, Catholic University of Valencia, Spain
Co-authors: Mariam Camarero, Cecilio Tamarit

A fully-fledged State-Space frame is developed for modeling panel time series that extends the simple framework generally employed into a panel-data time-varying parameters framework combining both fixed and varying components. Fixed parameters can be modeled either as country-specific or as a common parameter for the whole sample. Under determinate circumstances, this setting can be interpreted as a mean-reverting panel time-series model. Additionally, the mean fixed parameter can eventually include a deterministic trend. Regarding the equations governing the transition of the unobserved components, our structure allows for the estimation of different autoregressive alternatives that eventually include control instruments. The coefficients for the control instruments can be set up either as common for the panel, or as country-specific. The latter is particularly interesting for detecting asymmetries among individuals (countries) to common shocks. The code, that has been performed by the authors in Aptech Gauss, also allows for imposing restrictions regarding the relative size of variances of the error terms of the transition and measurement equations. Finally, we perform an empirical application of the proposed frame to estimate the estimation of a time-varying specification of the Okun's Law for a panel including both European and non-European industrialized countries on the support of its usefulness in solving complexities in macroeconomic empirical research.

CO322 Room Virtual R40 HIGHFREQUENCY Chair: Onno Kleen

C0393: **Sluggish news reactions: A combinatorial approach for synchronizing stock jumps**
Presenter: **Nabil Bouamara**, KU Leuven, Belgium
Co-authors: Kris Boudt, Sebastien Laurent, Christopher Neely

Sluggish news reactions manifest as gradual jumps and jump delays. In a panel of high-frequency intraday stock returns, these noisy jumps show up as a "sluggish cojump", i.e. jumps observed at close but distinct points in time. We synchronize these scattered jumps to recover the true common jump component. We apply our methods to investigate the local behavior of Dow 30 stock jumps in a short event window around an ETF jump.

C0848: **Probability distributions for realized covariance matrices**
Presenter: **Michael Stollenwerk**, Heidelberg University, Germany

Realized covariance matrices (RCs) are an important input to assess the risks involved in different investment allocations and it is thus useful to model and forecast them. To this end, generalized autoregressive score (GAS) models are employed. These models are ideal for comparing different probability distributions in terms of their ability to model and forecast RCs, since the dynamic parameters of the conditional observation density are updated by incorporating the shape of the distribution itself (via the scaled score of the log-likelihood). A novel type of probability distribution is derived and compared to all other probability distributions so far applied to RCs in the literature. The necessary inputs for the GAS models (Fisher information matrix and score) are derived for all distributions. An in-sample fit comparison confirms previous results that "fat-tailed" distributions outperform others and shows that the novel distribution achieves the best fit. Out-of-sample forecasting comparisons will be done using different economically relevant loss functions.

C1014: **Forecasting realized correlations: A MIDAS approach**
Presenter: **Anastasija Tetereva**, Erasmus University Rotterdam, Netherlands

Mixed data sampling (MIDAS) regression has received much attention in relation to modeling financial time series due to its flexibility. Previous work has mainly focused on forecasting realized volatilities and has rarely been used to predict realized correlations. The aim is to consider a MIDAS approach to forecast realized correlation matrices. A MIDAS model is estimated via nonlinear least squares (NLS) using an analytical gradient-based optimization. Based on the model confidence set (MCS) procedure we discover that the introduced approach is superior compared to the established heterogeneous autoregressive (HAR) model in terms of out-of-sample forecasting accuracy. This preeminence is due to the flexible data-driven origin of the MIDAS model. The latter results in higher economic value with regard to portfolio management applications. The improvement is considerable for longer forecasting horizons both in calm times and during periods of market turbulence.

C0469: **Beta-adjusted covariance estimation**
Presenter: **Kris Boudt**, UGent, VUB, VUA, Belgium
Co-authors: Kirill Dragun, Orimar Sauri, Steven Vanduffel

The increase in trading frequency of Exchanged Traded Funds (ETFs) presents a positive externality for financial risk management when the price of the ETF is available at a higher frequency than the price of the component stocks. The positive spillover consists in improving the accuracy of pre-estimators of the integrated covariance of the stocks included in the ETF. The proposed Beta Adjusted Covariance (BAC) equals the pre-estimator plus a minimal adjustment matrix such that the covariance-implied stock-ETF beta equals a target beta. We focus on a previous pre-estimator and derive the asymptotic distribution of its implied stock-ETF beta. The simulation study confirms that the accuracy gains are substantial in all cases considered. In the empirical part, we show the gains in tracking error efficiency when using the BAC adjustment for constructing portfolios that replicate a broad index using a subset of stocks.

CC867 Room K2.41 (Hybrid 09) CONTRIBUTIONS IN FORECASTING (HYBRID) Chair: Richard McGee

C1768: **Can we forecast better in periods of low uncertainty: The role of technical indicators**
Presenter: **Sam Pybis**, Manchester Metropolitan Universty, United Kingdom
Co-authors: Michalis Stamatogiannis, Olan Henry, Maria Ferrer Fernandez

The aim is to examine the importance of periods of high versus low financial uncertainty when forecasting stock market returns with technical predictors. The results suggest that technical predictors perform better in periods of low financial uncertainty and should be avoided due to poor forecasting performance in periods of heightened uncertainty. In-sample, we report disentangled R^2 statistics, and out-of-sample we show these results continue when forecasting the equity risk premium. We show similar results when forecasting the volatility of returns with technical predictors. We measure periods of heightened and low financial uncertainty in a regime-switching framework. Overall, our results provide insight into the mechanism that suggests that, when uncertainty rises, investors' opinions polarize leading to a breakdown of predictability based on technical indicators.

C1608: **Option-implied physical probabilities**
Presenter: **Richard McGee**, University College Dublin, Ireland
Co-authors: Thierry Post, Valerio Poti

A new analytical framework for forecasting the probability distribution of equity index returns is presented and applied using index option prices in an imperfect and incomplete market. The implied probabilities enhance a range of standard density forecasts in terms of out-of-sample predictive ability by including forward-looking pricing information. The economic significance of the improved forecasting strength manifests itself in improvements for Value-at-Risk calculations for index investors and delta hedging by market makers.

C1556: **Probabilistic forecasting with machine learning and big data**
Presenter: **Lubos Hanus**, UTIA AV CR, v.v.i, Czech Republic
Co-authors: Jozef Barunik

A distributional deep learning approach is proposed for probabilistic forecasting of economic time series. Being able to learn complex patterns from a large amount of data, deep learning methods are useful for decision making that depends on the uncertainty of a possibly large number of economic outcomes. Such predictions are also informative to decision-makers facing asymmetric dependence of their loss on outcomes from possibly non-Gaussian and non-linear variables. We show the usefulness of the approach on the three distinct problems. First, we use deep learning to construct data-driven macroeconomic fan charts that reflect the information contained by a large number of variables. Second, we obtain uncertainty forecasts of irregular traffic data. Third, we illustrate gains in the prediction of stock return distributions that are heavy-tailed and suffer from low signal-to-noise ratios.

C1662: **On the ability of risk-neutral densities to forecast the direction of change**
Presenter: **Maria Magdalena Vich Llompart**, Washington College, United States
Co-authors: Antoni Vaello Sebastia

Option-implied measures have been proven to be better forecasters of future realized measures than their historical counterparts. However, when it comes to the forecasting ability of option-implied Risk-Neutral densities (RND) there is no consensus in the literature. Therefore, being the prediction of a point estimate challenging, the focus is on the ability of the option-implied Risk-Neutral densities in predicting the future direction of the market instead. Risk-Neutral distributions are estimated from information extracted from option prices (for a 30-days horizon) using non-parametric techniques. Using a probit regression analysis, the aim is to study whether a daily change in the probability of a positive return estimated by the option-implied RNDs helps predict the sign of the future return 1, 2, 10 and 30 days ahead. We provide evidence for 11 international financial markets and we find that, overall, results are positive and significant.

CG031 Room Virtual R20 CONTRIBUTIONS IN APPLIED FINANCIAL ECONOMETRICS Chair: Kurt Lunsford

C0193: **Advance layoff notices and aggregate job loss**
Presenter: **Kurt Lunsford**, Federal Reserve Bank of Cleveland, United States

Establishment-level data are collected from advance layoff notices filed under the Worker Adjustment and Retraining Notification (WARN) Act since January 1990. First, we aggregate this data into a monthly, state-level, unbalanced panel and show that WARN layoffs lead state-level initial unemployment insurance (UI) claims, changes in the unemployment rate, and changes in private employment. Next, we use a dynamic factor model to aggregate the unbalanced state-level panel into a national WARN layoff index. This index moves closely with layoffs from Mass Layoff Statistics and the Job Openings and Labor Turnover Survey, but is timelier and covers a longer sample. Using a VAR, we show that an increase in the WARN layoff index generates increases in initial UI claims, and the job separation and unemployment rates. Finally, we show that the WARN layoff index improves pseudo-real-time forecasts of the unemployment rate.

C1269: **Stability of fundamental parity conditions in international finance**
Presenter: **Markus Moessler**, University of Hohenheim, Germany

The long-term stability of fundamental parity conditions in international finance is analyzed, in particular, the covered interest rate parity (CIP) condition. The CIP condition is based on the law of one price (LOP) on international financial markets and forms the basis for other fundamental parity conditions such as the uncovered interest rate parity (UIP) condition and the unbiased forward rate hypothesis. Recent studies found that the USD CIP condition held up well before but broke down after the Global Financial Crisis (GFC) of 2007-2008 and point to macroeconomic as well as financial explanations for this phenomenon. We propose to analyse the CIP condition in a vector error correction model (VECM) framework, which allows us to examine the stability of various relationships and trends driving international financial markets jointly. Moreover, whereas most studies focused on USD relationships we extend the analysis beyond USD relationships. This allows us to compare the effects of global factors, e.g., international financial stability, as well as local factors, e.g., national monetary policy, on fundamental parity conditions in international finance. In general, we confirm the result, that fundamental relationships in the international financial market changed after the GCF. However, no single cause can be identified, changes and potential explanations vary across currency pairs.

C0232: **Mind the Basel gap**
Presenter: **Matthijs Lof**, Aalto University, Finland
Co-authors: Petri Jylha

The Basel credit gap, the difference between a country's credit-to-GDP ratio and its estimated long-term trend, is used as a basis for setting the countercyclical regulatory capital buffers under the Basel III regulatory framework. Using international data from the BIS, we show that the Basel credit gap, estimated by a one-sided HP filter, is nearly equivalent to a naive 16-quarter change in the credit-to-GDP ratio and performs equally well in terms of predicting banking crises. We demonstrate that the near-equivalence between deviations from trend and simple changes occurs when the one-sided HP filter is applied to an I(1) process.

C0737: **Co-movement between commodity and equity markets revisited: An application of the Thick Pen method**
Presenter: **Sania Wadud**, The University of Aberdeen, United Kingdom
Co-authors: Marc Gronwald, Robert B Durand, Seungho Lee

The purpose is to analyse interdependence between the returns of a number of energy and non-energy commodities on the one hand and equities on the other based on Thick Pen Transform (TPT) methods: (i) Thick Pen Measure of Association (TPMA) and (ii) Multi-Thickness Thick Pen Measure of Association (MTTPMA). These metrics can be used to capture time-varying co-movement and co-movement across different time scales: this facilitates the analysis of the short-term and long-term features of the time series using both stationary/non-stationary data. Among the key findings is that, when considering long-term co-movement, energy index futures show an increase in co-movement with equities since the beginning of the financialisation period. There are asymmetric effects in cross-scale co-movement between various commodities and equities. The weak co-movement between equity and off-index futures, livestock and soybean-based commodities indicates diversification benefits for both short-term and long-term investors.

Monday 20.12.2021 08:15 - 09:30 Parallel Session M – CFE-CMStatistics

EO126 Room Virtual R20 THEORY AND COMPUTATION IN INFERENCE FOR STOCHASTIC PROCESSES Chair: Hiroki Masuda

E0451: **Hawkes processes with a state-dependent factor**
Presenter: **Ioane Muni Toke**, CentraleSupelec, France
A point process model is proposed for order flows in limit order books, in which the conditional intensity is the product of a Hawkes component and a state-dependent factor. In the LOB context, state observations may include the observed imbalance or the observed spread. We provide full technical details for the computationally efficient estimation of such processes, using either direct likelihood maximization or EM-type estimation. Applications include models for bid and ask market orders, or for upwards and downwards price movements. Empirical results on multiple stocks traded in Euronext Paris underline the benefits of state-dependent formulations for LOB modeling, e.g. in terms of goodness-of-fit to financial data.

E0982: **Scaling limit of some piecewise deterministic Markov processes via multi scale analysis**
Presenter: **Kengo Kamatani**, ISM, Japan
Recently, piecewise deterministic Markov processes have gained interest in the Monte Carlo community in the context of scalable Monte Carlo integration methods. We will discuss high-dimensional scaling limits for some piecewise deterministic Markov processes. We will describe these results using multiscale analysis, which is a useful technique for this purpose. We will also highlight two types of scaling limits corresponding to the tangential direction and the linear direction to the log-density contour.

E1403: **Nonparametric inference of coefficients of self-exciting jump-diffusion processes**
Presenter: **Arnaud Gloter**, Universite d Evry Val d Essonne, France
Co-authors: Chiara Amorino, Charlotte Dion, Sarah Sarah Lemler
A one-dimensional diffusion process with jumps driven by a Hawkes process is considered. We are interested in the estimations of the volatility function and of the jump function from discrete high-frequency observations in a long time horizon. We first propose to estimate the volatility coefficient. For that, we introduce a truncation function in our estimation procedure that allows us to take into account the jumps of the process and estimate the volatility function on a linear subspace of $L^2(A)$ where A is a compact interval of $\mathbb{R}$. We obtain a bound for the empirical risk of the volatility estimator and establish an oracle inequality for the adaptive estimator to measure the performance of the procedure. Then, we propose an estimator of a sum between the volatility and the jump coefficient modified with the conditional expectation of the intensity of the jumps. We also establish a bound for the empirical risk for the non-adaptive estimators of this sum and an oracle inequality for the final adaptive estimator.

EO090 Room Virtual R21 STATISTICAL MODELS FOR SURVIVAL DATA I Chair: Marialuisa Restaino

E0218: **A new cut-point PH-distribution to fit a Survival data set**
Presenter: **Christian Acal**, University of Granada, Spain
Co-authors: Juan Eloy Ruiz-Castro
Phase-type distributions (PHD) are a suitable candidate to model complex problems in an algorithmic and computational way. Among other properties, PHD class stands out for being dense in the set of probability distributions on the non-negative half-line, which enables to approximate as much as desired any non-negative probability distribution. Nevertheless, the drawback is that the estimation of the PHD parameters is not a simple task since the PHD representation is not unique. Although the fitting is really acceptable on most occasions, two aspects raise some concern in the optimization problem: the number of parameters to be estimated is usually high, and at times PHDs do not provide good results in the tails of the distribution. A novel methodology based on PH distributions with multiple cut-points is proposed. In particular, a new distribution called multiple cut-points PHD is introduced to solve the lack of power in the distribution tails and to reduce the number of parameters in the estimation. This class of distributions is studied in detail and several associated measures are worked out. An EM algorithm is developed for parameter estimation. The results have been implemented in R-cran and multiple examples are developed.

E0530: **The underrecognized potential of logistic regression to analyze survival data from clinical trials**
Presenter: **Paul Blanche**, University of Copenhagen, Denmark
Co-authors: Thomas Scheike
Guidelines from regulatory agencies have long emphasized the benefit of adjusting for baselines variables in the analysis of randomized clinical trials. Recently, they put forward the specific use of G-computation based on logistic regression for analyzing binary outcomes. This approach is more powerful than an unadjusted analysis and it provides a robust estimator of the average treatment effect which is consistent under arbitrary model misspecification. We argue that when the main outcome of a clinical trial is t-year survival or similar, a similar approach can be used, even when lost-of follow-up occurs and creates right-censored data. We show that under mild conditions, the inverse probability of censoring weighting provides similar robust inference as in the case of binary outcomes. As compared to alternatives based on hazards regression, the estimand of this approach is clearly defined and does not rely on arbitrary model assumptions. The approach is straightforward to implement with standard software and provides a simple and transparent approach to leverage the information in baseline variables.

E0871: **New tools for analyzing doubly truncated data**
Presenter: **Carla Moreira**, University of Minho, Portugal
Co-authors: Jacobo de Una-Alvarez, Rosa Crujeiras
The analysis of doubly truncated data is relevant in epidemiological applications, when the observation of the lifetime of interest is limited to events between two specific calendar dates. This implies that small or large times are less probably observed and thus properly corrections to the estimators must be done in order to avoid biased estimations that may lead to wrong conclusions. Given the aforementioned motivation, the interest of the scientific community in the phenomenon of random double truncation has significantly grown, particularly in fields like Epidemiology and Survival Analysis and this has motivated the development of software routines that could facilitate a proper analysis of this kind of data. DTDA, built in 2010, was the first R library spreading the methods for the analysis of doubly truncated data. Due the constant challenge to develop new methods devoted to double truncation, the need to follow the new statistical methods with software development is pressing. In addition to the implemented algorithms to estimate the cumulative distribution function, the renewed DTDA package outfits smoothing methods to estimate the kernel density function and hazard function, including the bandwidth selection procedures for the density function. Different real datasets from different areas were also included.

EO517 Room Virtual R22 BRANCHING PROCESSES: THEORY, COMPUTATION AND APPLICATIONS I Chair: Ines M del Puerto

E1517: **Stochastic evolution of particle systems for branching random walks in non homogeneous and homogeneous environments**
Presenter: **Elena Yarovaya**, Lomonosov Moscow State University, Russia
For the study of stochastic evolution of particle systems, an approach is applied which is focused on continuous-time symmetric branching random walks on multidimensional lattices. The main object of study is the limit distribution of particles on the lattice and their moments. The limit theorems on the asymptotic behavior of the Green function for transition probabilities were established for random walks under different assumptions

on a variance of random walk jumps. For supercritical branching random walks with one initial particle and a finite number of particle generation centers called branching sources and located at the lattice points, it is shown that the amount of positive eigenvalues of the evolutionary operator of the mean number of particles, counting their multiplicity, does not exceed the amount of branching sources with a positive intensity on the lattice. We demonstrate that the appearance of multiple lower eigenvalues in the spectrum of the evolutionary operator can be caused by a kind of 'symmetry' in the spatial configuration of branching sources. If initially there is one particle at each lattice point which can walk over the lattice and produce offspring at every lattice point by a critical Markov branching process under the assumption that the underlying walk is recurrent, the convergence of the distribution of the particle field to the limit stationary distribution is obtained.

E1478: **Estimation of the carrying capacity in branching processes via weighted least square estimation**
Presenter: **Carmen Minuesa Abril**, University of Extremadura, Spain
Co-authors: Peter Braunsteins, Sophie Hautphenne

In the context of branching processes that describe populations with logistic growth, the purpose is the estimation of their parameters. We focus on the parameter representing the maximum population size of the species that an environment can maintain in view of the available resources, known as the carrying capacity. To this end, we make use of the family of discrete-time population-size-dependent branching processes. In these processes, individuals reproduce independently with an offspring distribution that depends on the current number of individuals in the population. Assuming a parametric framework for the probability laws describing the reproduction, we present some weighted least square estimators of the target parameters. The asymptotic properties, such as consistency and asymptotic normality, of the proposed estimators are illustrated via several examples. We also apply them to estimate the carrying capacity of an endangered population of birds: the black robin.

E1519: **Comparing numerical results for branching random walks in non random and random media**
Presenter: **Vladimir Kutsenko**, Lomonosov Moscow State University, Russia
Co-authors: Elena Yarovaya

Continuous-time branching random walks (BRWs) on a multidimensional lattice in a random branching medium are considered. The branching medium may contain a finite or non-finite number of particle generation sources. The underlying walk of particles is symmetric, homogeneous by space, and irreducible with a finite variance of jumps. The vast majority of the results obtained in the theory of BRWs in random media are asymptotic. In such BRWs, at large times, rare fluctuations of the medium may lead to anomalous properties of a particle field such as "intermittency". At the same time, the study of BRWs at finite time intervals seems to be a difficult task that has not yet been solved satisfactorily enough. Thus, the main goal of the simulation is to investigate whether it is possible to obtain qualitative and quantitative results predicted by the theory already at finite times. A similar task for BRWs in non-random media has been previously considered. However, to the best of our knowledge, there are no similar studies related to the simulation of BRWs in random branching media. Here, the phenomenon of intermittency is of the most significant interest and importance. Based on the simulation results, we managed to show that intermittency can be observed in random media even over finite time intervals.

EO555 Room Virtual R23 BELIEFS, RISK AND UNCERTAINTY IN ARTIFICIAL INTELLIGENCE II Chair: Davide Petturiti

E0529: **Markovian dynamics under Wasserstein uncertainty**
Presenter: **Max Nendel**, Bielefeld University, Germany

A class of time-homogeneous continuous-time Markov processes is considered with transition probabilities bearing a nonparametric uncertainty. The uncertainty is modelled by considering perturbations of the transition probabilities within proximity in Wasserstein distance. As a limit over progressively finer time periods, on which the level of uncertainty scales proportionally, we obtain a convex semigroup (family of convex transition operators) satisfying a nonlinear PDE in a viscosity sense. A remarkable observation is that, in standard situations, the nonlinear transition operators arising from nonparametric uncertainty coincide with the ones related to parametric drift uncertainty. On the level of the generator, the uncertainty is reflected as an additive perturbation in terms of a convex functional of first-order derivatives. We additionally provide sensitivity bounds for the convex semigroup relative to the reference model.

E0760: **Generalised measurement error models**
Presenter: **Thomas Augustin**, LMU Munich, Germany

From social sciences and econometrics to biometrics and technometrics, measurement error is an omnipresent problem: many variables of interest are latent and prone to non-negligible measurement error. A bundle of methods has been developed in the literature to adjust inference for error-proneness of data. However, these methods typically rely on the so-called classical error model/testing theory, making rather strong, and often untestable, assumptions about the eo ipso unobservable error process. However, in the spirit of Manski's Law of Decreasing Credibility, untenable assumptions undermine the practical relevance of the results. Against this background, generalised measurement error models are proposed. The measurement error distribution is flexibly described by a credal set of probability distributions, for instance, by a neighbourhood model or by a parametrically constructed model based on a set of parameter values. Nakamura's method of corrected score functions will be extended to deal with set-valued expectations, enabling the derivation of set-valued estimators that satisfy some generalised criteria of unbiasedness. Finally, some ideas are discussed on how to overcome in addition some doubtful independence assumptions in the measurement error modelling process.

E0763: **Aggregation of comonotone risks using robust distribution functions**
Presenter: **Ignacio Montes**, University of Oviedo, Spain

Comonotonicity is a dependence structure modelling random variables that increase or decrease simultaneously. Comonotone random variables possess not only interesting mathematical properties, but they also appear in several applications, mainly in decision theory and finance. In the latter field, understanding random variables as risks, there are three well-known properties related to comonotonicity: (i) there exists a known expression for the cdf of the aggregated risks under comonotonicity; (ii) the aggregation of the risks assuming comonotonicity is a conservative approximation of the aggregated risk with respect to the stop-loss premium; and (iii) the VaR of the aggregation of comonotone risks can be decomposed as the aggregation of the VaR of the marginal risks. A natural question appears: what happens when the probability distribution of the risks is only partially known? In those cases, the cdfs can be replaced by robust cdfs (known as p-boxes in the imprecise probability literature) giving lower and upper bounds to the values of the real but unknown cdfs. The three aforementioned properties of comonotone risks are shown still hold when using robust cdfs, and proves that one robust model performing particularly well in this setting is the Kolmogorov model.

EO246 Room Virtual R24 SOME RECENT RESULTS ON STATISTICAL MODELLING Chair: Geoffrey McLachlan

E1508: **Finite sample inference for nonlinear autoregressive models**
Presenter: **Hien Nguyen**, University of Queensland, Australia

Autoregressive processes are fundamental modelling tools in time series analysis. To conduct inference for such models usually requires asymptotic limit theorems. We establish finite sample-valid tools for hypothesis testing and confidence set construction in such settings. These constructions can be used to conduct inference in highly nonlinear and irregular scenarios.

E1497: **A selective survey of mixture models with flexible distributions**
Presenter: **Sharon Lee**, University of Queensland, Australia

Extensive growth in the literature on multivariate flexible distributions in the past few decades has provided an abundance of choices for mod-

elling the distribution of non-normal data. Some of these have been applied to mixture models and cluster analyses, in particular those that can accommodate skewness and heavy-tailedness. We provide an updated survey on recent developments in this area, focusing on the more popular proposals and those with publicly available software implementations. The different formulations used to construct these distributions may render them suitable for different applications. Their performances at modelling different types of asymmetric data are studied using simulations.

E1529: **Statistical modelling with noisy labels: An application to fraud detection**
Presenter: **Daniel Ahfock**, University of Queensland, Australia
Co-authors: Min Zhu

Label noise is a practical consideration in many applications of supervised learning. Ground-truth labels may not be readily obtainable due to the high cost of acquisition. This limitation is encountered in accounting fraud detection, where regulators do not have the resources to investigate every firm for evidence of fraud. Some instances of fraud in the population may go undetected. There has been growing interest in the development of classifiers for the detection of accounting fraud given financial data from accounting reports. The training data consists of yearly financial data from each firm, and whether the firm was cited for accounting fraud by a regulator. We consider the development of a statistical model for accounting fraud detection, allowing for label noise in the training set. We propose to treat the ground-truth fraud status labels as latent variables and to model the label noise process. Maximum likelihood estimation can be carried out using the expectation-maximisation algorithm. We show how our approach can be used to train generalised additive models given noisy labels. We present an application of the method to historical fraud detection data.

EO358 Room Virtual R26 DATA SCIENCE AND CYBERSECURITY Chair: Clara Grazian

E0428: **Cluster analysis of hacker groups from terminal commands issued in honeypots**
Presenter: **Nicholas Heard**, Imperial College London, United Kingdom

In computer security, a honeypot is a host located within a computer network designed to entice malicious intruders. From interactive sessions initiated by users engaging with a honeypot, we are able to harvest the commands they issued as well as other information about the session such as timings, operating system and IP address. These session commands provide a rare insight into the operational modes of cyber attackers, such as their automated or interactive nature, the individual scripting styles and their overall objectives. The volume of traffic passing through a honeypot can be surprisingly high, and so automating the understanding of these sessions, classifying them and detecting new emerging styles provides a challenging data science research problem.

E0636: **Assessing the invertibility of deep biometric representation**
Presenter: **Yiwei Wang**, University of New South Wales, Australia
Co-authors: Clara Grazian

Biometric recognition is increasingly applied in practice given the worries about security problems. Although various approaches have been used to protect databases of biometric systems, there still exists the possibility that the template of a person is stolen. Since biometric traits are physically connected to a person, we focus on the ability of networks to avoid the attach of using inverse imagining from the deep representation. We use convolutional neural networks to generate the deep representations and a GAN to reconstruct the image. Then we analyse the invertibility of the representation through random forest regression. We try to understand what features of the NN have the largest impact on the representation invertibility.

EO380 Room Virtual R28 FUNCTIONAL AND HIGH-DIMENSIONAL DATA ANALYSIS Chair: Jeng-Min Chiou

E1408: **Functional survival analysis with convex clustering**
Presenter: **Yuko Araki**, Shizuoka University, Japan

The survival analysis which contains several classes of functional process, measured at baseline, is investigated. In our method, first, we identify the class of individual trajectory by the proposed functional convex clustering method, and as a second stage, class information is used in the Cox proportional hazards regression models to assess the risk of mortality during the follow-up period. We assess the performance of the proposed model in a simulation study. For illustration, the proposed method was applied to the cohort study in Japan.

E1748: **Bayesian deep neural networks: Optimality and adaptivity**
Presenter: **Yongdai Kim**, Seoul National University, Korea, South

A Bayesian model for learning a certain sparse deep neural network is considered and an efficient MCMC algorithm is developed. In addition, we derive the posterior contraction rate for the proposed Bayesian model which is minimax optimal for various nonparametric regression models. Moreover, we prove that this optimality is adaptive to the unknown smoothness of the true function. By analyzing several benchmark data with our Bayesian model, we illustrate that the Bayesian model is superior to other nonparametric estimators.

E1243: **Truncated estimation for varying-coefficient functional linear models**
Presenter: **Hidetoshi Matsui**, Shiga University, Japan

The focus is on the problem of estimating a varying-coefficient functional linear model, where the predictor is a function of time and the scalar response depends on not only a functional predictor but also an exogenous variable. The aim is to estimate the model so that the functional predictor does not relate to the response after a certain point in time at any value of the exogenous variable. To do so, we apply the sparse regularization to shrink the corresponding domain of the coefficient function towards exactly zero. Simulation studies are conducted to investigate the effectiveness of the proposed method, and we also apply the method to the analysis of crop yield data.

EO802 Room Virtual R34 BAYESIAN METHODS FOR EXTREME EVENTS Chair: Isadora Antoniano-Villalobos

E0755: **A Bayesian framework for Poisson process characterization of extremes with uninformative prior**
Presenter: **Theo Moins**, Inria, France
Co-authors: Stephane Girard, Julyan Arbel, Anne Dutfoy

Combining extreme-value theory with Bayesian methods offers several advantages, such as the availability of posterior predictive inference or the ability to study irregular cases for frequentist statistics. When no prior information is available, objective Bayes aims at using an external rule to construct a prior distribution. In particular, we focus on the use of Jeffreys prior for the Poisson process characterization of extremes, a model which generalizes the two most frequent ones i.e. the extreme-value distribution (EVD) for block-maxima and the generalized Pareto distribution (GPD) for peaks-over-threshold. After showing posterior propriety results, we also compare different reparametrisations of the Poisson process to facilitate sampling by Markov chain Monte Carlo (MCMC). In particular, the influence of a hyperparameter of the model and the interest of parameters orthogonality for Bayesian inference are investigated on simulated data.

E0785: **Bayesian non-asymptotic extreme value hierarchical models**
Presenter: **Antonio Canale**, University of Padua, Italy
Co-authors: Enrico Zorzetto, Marco Marani

A general Bayesian hierarchical model is introduced for estimating the probability distribution of extreme values of intermittent random sequences. Our approach avoids the asymptotic assumption typical of the traditional extreme value theory, and accounts for the possible underlying variability

in the distribution of event magnitudes and occurrences, which are described through a latent temporal process. Focusing on daily rainfall extremes, the structure of the proposed model lends itself to incorporating a prior geo-physical understanding of the rainfall process. Empirical performance is illustrated showing less tendency to overfitting and better out-of-sample predictions. Spatio-temporal extensions of the model are also discussed.

E1028: **Spatiotemporal wildfire modeling through point processes with moderate and extreme marks**
Presenter: **Jonathan Koh**, EPFL, Switzerland
Co-authors: Thomas Opitz

Accurate spatiotemporal modeling of conditions leading to moderate and large wildfires provides a better understanding of mechanisms driving fire-prone ecosystems and improves risk management. We here develop a joint model for the occurrence intensity and the wildfire size distribution by combining extreme-value theory and point processes within a novel Bayesian hierarchical model, and use it to study daily summer wildfire data for the French Mediterranean basin during 1995–2018. The occurrence component models wildfire ignitions as a spatiotemporal log-Gaussian Cox process. Burnt areas are numerical marks attached to points and are considered extreme if they exceed a high threshold. The size component is a two-component mixture varying in space and time that jointly models moderate and extreme fires. We capture the non-linear influence of covariates (Fire Weather Index, forest cover) through component-specific smooth functions, which may vary with season. We propose estimating shared random effects between model components to reveal and interpret common drivers of different aspects of wildfire activity. This leads to increased parsimony and reduced estimation uncertainty with better predictions. Specific stratified subsampling of zero counts is implemented to cope with large observation vectors. We compare and validate models through predictive scores and visual diagnostics.

EC853 Room K E. Safra (Multi-use 01) MULTIVARIATE AND HIGH-DIMENSIONAL STATISTICS (IN-PERSON) Chair: Ioulia Papageorgiou

E1604: **Multivariate outlier detection with ICS and application to statistical quality control for autocorrelated data**
Presenter: **Ioulia Papageorgiou**, Athens University of Economics and Business, Greece
Co-authors: Stefanos Voutsinas

Detection of special structures in the data, e.g. outliers, is an issue of high priority in Statistical Quality Control (SQC) because it can be an indicator of an out-of-control production line. The early detection of such measurements is essential and this problem is challenging when the population of interest is multivariate. Multivariate data on the other hand is more often the case than the exception in nowadays applications. Another parameter we consider for the SQC application is the autocorrelation among observations which is again a realistic scenario in this field. Most of the existing methodologies for detecting outliers fail to reveal those measurements for both multivariate case and autocorrelation. The aim is to examine the use of Invariant Coordinate Selection (ICS) in Statistical Quality Control and especially in detecting extreme measurements in case of correlated multivariate data. The experiments include various choices of scatter pairs for the use of ICS, the degree of correlation and the mechanism of generating the outliers. The performance of ICS method for detecting outliers in SQC is compared with the Mahalanobis distance and T^2 Hotelling chart plot. The comparison and evaluation are based on (i) the percentages of correct True Positive (TP) detection of outliers and false detection, False Positives (FP), events.

E1565: **Extreme partial least-squares regression**
Presenter: **Meryem Bousebata**, Inria, France
Co-authors: Stephane Girard, Geoffroy Enjolras

A new approach, called Extreme-PLS, is proposed for dimension reduction in regression and adapted to distribution tails. The goal is to find linear combinations of predictors that best explain the extreme values of the response variable by maximizing the associated covariance. This adaptation of the PLS estimator to the extreme-value framework is achieved in the context of a non-linear inverse regression model. In practice, it allows quantifying the effect of the covariates on the extreme values of the response variable in a simple and interpretable way. Moreover, it should yield improved results for most estimators dealing with conditional extreme values thanks to the dimension reduction achieved in the projection step. From the theoretical point of view, the asymptotic normality of the Extreme-PLS estimator is established under a heavy tail assumption but without recourse to linearity or independence assumptions. The performance of the method is assessed on simulated data. Finally, the Extreme-PLS approach is used to analyze the influence of various parameters on extreme cereal yields collected on French farms.

E1312: **High-dimensional changepoint selection with fused graphical models**
Presenter: **Alex Gibberd**, Lancaster University, United Kingdom
Co-authors: Sandipan Roy

The estimation of graphical models which have piece-wise constant support in terms of a time-evolving precision matrix is considered. Utilising a regularised likelihood framework we consider the impact of simultaneous smoothing and shrinkage penalties on the precision matrix. This M-estimator is then used to construct a gain-function which we use to search for changepoints. We derive bounds on both the ability of our estimator to recover the correct precision matrix entries, edge structure in the graphical model, and associated changepoints. We also illustrate how the method can be used in practice to segment complex high-dimensional data.

EG017 Room K0.19 (Hybrid 04) CONTRIBUTIONS IN TIME-VARYING APPROACHES Chair: Tommaso Proietti

E1212: **Time-varying GMM estimation in structural time series models**
Presenter: **Yu Bai**, Bocconi University, Italy

The aim is to develop time-varying continuously updated GMM estimation and inferential theory for moment conditional models whose coefficients vary stochastically and smoothly over time. We propose two new tests of structural stability in this context. After deriving the asymptotic properties of the estimators and test statistics, we assess finite sample performance by an extensive Monte-Carlo study and illustrate their application by an empirical example on conditional asset pricing models with stochastic discount factor representation.

E1717: **Time-varying instrumental variable estimation: A bootstrap approach**
Presenter: **Charisios Grivas**, Birkbeck University, United Kingdom

A bootstrap approach is considered for a Hausman-type test statistic for econometric models with endogenous regressors whose coefficients are allowed to vary over time both deterministically or stochastically. We compare the finite sample performance of the asymptotic and the bootstrap version of the test by means of Monte Carlo simulations. The bootstrap test statistic appears to have the proper size and higher power. More importantly, since these tests rely on a bandwidth parameter, it is shown that the size and the power of the bootstrap test are invariant to the choice of the bandwidth parameters.

E1241: **Time-varying co-movements among financial and financialised assets: The cyclical variation of the cross-asset nexus**
Presenter: **Jiaying Wu**, Brunel University, United Kingdom
Co-authors: Menelaos Karanasos, Starvoula Yfanti

The phenomenal financialisation of non-financial markets over the past two decades can act as a potent threat to financial stability given its contribution to financial contagion and systemic risk build-ups. We study the dynamic interdependence between stocks, a risky and financial by definition asset class, and the financialised assets from the real estate and commodity markets. Through a trivariate DCC-MIDAS setting, we analyse short- and long-run time-varying correlation dynamics among global stocks, real estate benchmarks, and five different commodity types: energy, precious metals, industrial metals, agriculture, and livestock. The empirical results demonstrate either strong cross-asset interlinkages

highly dependent on the state of the economy in many cases or weak connectedness for certain assets with hedging or safe-haven properties. We further investigate the macro-relevance and crisis-vulnerability of the evolution of the correlation by unveiling the macro-determinants of asset co-movements and the significant contagion effects during crisis periods. The economic environment plays a key role as a contagion transmitter, while the uncertainty channel intensifies the macro impact on the cross-asset nexus.

EG059 Room Virtual R25 CONTRIBUTIONS IN BAYESIAN STATISTICS II Chair: Tomonari Sei

E1617: **A correlation-shrinkage prior for the 2-dimensional Wishart model**
Presenter: **Tomonari Sei**, The University of Tokyo, Japan
Co-authors: Fumiyasu Komaki

For the two-dimensional Wishart model, we propose a scale-invariant and permutation-invariant prior distribution that shrinks the correlation coefficient. The prior is characterized by a uniform distribution for Fisher's z-transformation of the correlation coefficient. The Bayesian predictive density based on our prior is shown to be minimax under the Kullback-Leibler loss.

E1622: **Stochastic time-discrete SIR models and particle filtering**
Presenter: **Koichi Yano**, Komazawa University, Japan
Co-authors: Tae Okada

A novel method is proposed for estimating a stochastic time-discrete SIR (Susceptible, Infectious, or Recovered) model using particle filtering with time series of the number of infected and recovered persons. The time-discrete/time-continuous SIR model, which is one of the compartmental models of epidemiology, is a prevailing model in theoretical epidemiology. By combining the model with state and parameter estimation, the more precise forecasting of the number of infected persons is realized. My new method, which combines state estimation based on particle filtering and parameter estimation based on the Nelder-Mead method/the particle Metropolis-Hasting method, can reproduce the actual time series of the number of infected persons more faithfully. Using it, we estimate a stochastic time-discrete SIR model with the daily COVID-19 time series for Japan, simulate the number of infected persons in Japan, and forecast the number of infected people based on the estimated model. In addition, we analyze the factors behind the spread and decrease of COVID-19 infection in Japan based on the estimated states and parameters. This research is a great contribution to the global economy because the method will enable governments to quickly implement better infection control measures.

E1624: **Bayesian sparse seemingly unrelated regression model with variable and covariance selection**
Presenter: **Dongu Han**, Korea University, Korea, South
Co-authors: Taeryon Choi

Seemingly Unrelated Regression(SUR) is a general framework that can accommodate many useful models, such as multivariate regression or vector autoregressive model. With the era of big data, the number of predictors and the equations to be estimated simultaneously can be both large compared to the sample size. We handle this problem by adopting a variant of horseshoe prior to the parameters. Implementing this prior, we provide an efficient Markov Chain Monte Carlo(MCMC) algorithm without any additional tuning procedures. We also provide some theoretical results, indicating our model works well under mild conditions and provides better estimation compared to the conventional ones. Additionally, we also propose a variational Bayesian method that brings much computational gain without sacrificing the precision of estimation. Several empirical studies show that our proposed method works better than conventional ones.

CO780 Room Virtual R18 RECENT DEVELOPMENTS ON STATISTICAL LEARNING AND ITS APPLICATIONS Chair: Wei Zhou

C0909: **GSLM: Structure learning via unstructured kernel-based M-regression**
Presenter: **Yeheng Ge**, Shanghai University of Finance and Economics, China
Co-authors: Xingdong Feng, Xin He

In statistical learning, identifying underlying structures of true target functions based on observed data is crucial in facilitating subsequent modeling and analysis. Unlike most of those existing methods that focus on some specific settings under certain model assumptions, a general and novel framework is proposed for recovering true structures of target functions in a reproducing kernel Hilbert space (RKHS). The proposed framework is inspired by the fact that the gradient functions can be employed as a valid tool to learn underlying structures, including sparse learning, interaction selection and model identification. It is easy to be implemented by taking advantage of the nice properties of the RKHS. More importantly, it admits a wide range of loss functions, and thus includes many scenarios as its special cases, such as mean regression, quantile regression, likelihood-based classification, and margin-based classification, which is also computationally efficient by solving convex optimization tasks. The asymptotic results of the proposed framework are established within a rich family of loss functions without any explicit model specifications. The superior performance of the proposed framework is also supported by a variety of simulated examples and a real case study.

C1468: **Efficient learning of quadratic variance function directed acyclic graphs via topological layers**
Presenter: **Wei Zhou**, Xiamen University, City University of Hong Kong, China
Co-authors: Wei Zhong

Directed acyclic graph (DAG) models are widely used to represent causal relationships among random variables in many application domains. A special class of non-Gaussian DAG models is studied, where the conditional variance of each node given its parents is a quadratic function of its conditional mean. Such a class of non-Gaussian DAG models are fairly flexible and admit many popular distributions as special cases, including Poisson, Binomial, Geometric, Exponential, and Gamma. To facilitate learning, we introduce a novel concept of topological layers, and develop an efficient DAG learning algorithm. It first reconstructs the topological layers in a hierarchical fashion and then recoveries the directed edges between nodes in different layers, which requires much less computational cost than most existing algorithms. Its advantage is also demonstrated in a number of simulated examples, as well as its applications to two real-life datasets, including an NBA player statistics data and cosmetic sales data collected by Alibaba.

C1532: **Censored quantile regression based on multiply robust propensity scores**
Presenter: **Xiaorui Wang**, East China Normal University, China

Censored quantile regression (QR) has elicited extensive research interest in recent years. One class of methods is based on an informative subset of a sample, selected via the propensity score (PS). PS can either be estimated using parametric methods, which poses the risk of misspecification, or obtained using nonparametric approaches, which suffers from the curse of dimensionality. We propose a new estimation method based on multiply robust PS for censored QR. This method only requires one of the multiple candidate models for PS to be correctly specified, and thus, it provides a certain level of resistance to the misspecification of parametric models. Large sample properties, such as the consistency and asymptotic normality of the proposed estimator, are thoroughly investigated. Extensive simulation studies are conducted to assess the performance of the proposed estimator. The proposed method is also applied to a study on human immunodeficiency viruses.

CO220 Room Virtual R27 ECOSTA JOURNAL SESSION I Chair: Alessandra Amendola

C0219: **Nonparametric estimation for multivariate time-varying models: Theory and practice**
Presenter: **Yayi Yan**, Monash University, Australia
Co-authors: Jiti Gao, Bin Peng

Multivariate dynamic time series models are widely encountered in practical studies, e.g., modelling policy transmission mechanisms and measuring connectedness between economic agents. To better capture the dynamics, we initiate the study on a time-varying VMA infinity model, develop a time-varying counterpart of the conventional BN decomposition, and establish the inferential theories associated with the trending function. Then, by imposing a more detailed structure on the data generating process (i.e., the time-varying VAR(p) model), we establish theories for the impulse response functions, which are of great interest to describe how the economy reacts over time to economic shocks. The asymptotic results are established subject to both short-run timing and long-run restrictions, respectively. Third, we examine the theoretical results through extensive simulated and real data studies.

E1659: **Comparison of machine learning algorithms for an HAR problem using rotation times series data**
Presenter: **Raphael Brard**, University of Nantes, France
Co-authors: Lise Bellanger, Aymeric Stamm, Pierre Drouin, Laurent Chevreuil, Fanny DOISTAU

In the last decade, there has been a growing interest in human activity recognition (HAR) algorithms based on inertial sensor data. The company UmanIT and the Department of Mathematics Jean Leray in Nantes have developed a solution for facilitating gait analysis on straight walk phases. To facilitate its use in real-life situations, we implemented and compared different machine learning algorithms (Support Vector Machine, Decision Tree, k-nearest neighbors and logistic regression) associated with post-treatment aiming at reducing false positive detections from a human daily recording. The data was collected by a motion sensor in the form of a unit quaternion time-series recording the hip rotation over time. This time series was then transformed into a real-valued time series of geodesic distances between consecutive quaternions. Moving average and moving variance versions of this time series were fed to the machine learning algorithms in order to train, tune and test our models. To compare the different models, we used metrics to assess the classification performances (AUC, accuracy) as well as metrics to assess change point detection capability and computation time. SVM stood out in terms of performance while decision trees led to the best compromise between performances and computation time.

CO048 Room Virtual R29 SUSTAINABLE FINANCE I Chair: Monica Billio

C0392: **Transition factors and market-implied credit risk**
Presenter: **Michele Costola**, Ca' Foscari University of Venice, Italy
Co-authors: Katia Vozian

The energy transition to a low-carbon economy has a time horizon of c. 30 years (EU NetZero 2050). We aim to identify which environmental factors pertaining to the low-carbon transition of a firm relate to the firm market-implied credit risk and how. We construct four datasets merging CDS spreads with time horizon 1-5-10-30 years to E-transition factors and to traditional determinants of CDS spreads. The resulting samples cover over 200 non-financial European firms in the period from 2010 to 2020. We analyze in a panel regression the relation between credit risk, as proxied by CDS spread, and emissions data over short and long-term time horizons. Preliminary results show that higher GHG Emissions Scope 1 relate to higher credit risk, irrespective of the data provider, and the magnitude of the relation diminishes at time horizons of 10 and 30 years, although is still significant.

C0449: **Green sentiment, stock returns and corporate behavior**
Presenter: **Stefano Ramelli**, University of Zurich, Switzerland
Co-authors: Marie Briere

A new method is proposed to estimate non-fundamental demand for green financial assets based on the arbitrage activity of exchange-traded funds (ETFs). By estimating the monthly abnormal flows into environment-friendly ETFs, we construct a Green Sentiment Index capturing shifts in investors' appetite for environmental responsibility not yet priced in the value of the underlying assets. Our measure of green sentiment differs significantly from the climate-related news and attention indexes proposed by the extant literature, and it has additional explanatory power on both stock returns and corporate decisions. Over the period 2010-2020, changes in green sentiment anticipate a lasting stock out-performance by more environmentally responsible firms (of approximately 60 basis points over six months for a one-standard-deviation higher green sentiment), as well as an increase in their capital investments and cash holdings.

C0400: **ESG factors and firms' credit risk**
Presenter: **Vittoria Cerasi**, Bicocca University, Italy
Co-authors: Matteo Manera, Laura Bonacorsi

The link between the risk of default and Environmental, Social, Governance (ESG) factors is studied using supervised machine learning techniques on a cross-section of European listed companies. We focus on ESG factors instead of ESG scores, to avoid relying on unobserved models used by rating companies to construct these ratings. The advantage of this approach is also that our results can be applied to non-rated corporations. We proxy credit risk by using the Altman z-score, which is a linear combination of accounting ratios used to classify companies in different categories according to their risk of default. Our sample is a cross-section of 1251 European firms in the year 2019, and it includes 590 candidate variables. Due to the huge number of variables involved, we employ techniques of supervised machine learning, in particular, the Least Absolute Shrinkage and Selection Operator (LASSO), to select the relevant explanatory variables. Since our objective is to predict the sign of the selected variables on the risk of default, we use Lasso for inference methods. The preliminary results show that a selection of ESG factors in addition to the usual accounting ratios helps to explain a firm's probability of default. We also develop a model to explain short-term credit rationing augmented for ESG factors to interpret the results.

CO168 Room Virtual R30 ECONOMETRIC FORECASTING Chair: Robert Kunst

C1310: **Measuring uncertainty to identify financial instability**
Presenter: **Ines Fortin**, Institute for Advanced Studies, Austria
Co-authors: Jaroslava Hlouskova, Leopold Soegner

A new index measuring financial (in)stability for Austria and for the Euro area is introduced based on market data (equity, bond, money, and foreign exchange markets). The new index is a so-called uncertainty index, which follows the existing methodology and is fundamentally different from the well-established financial stability indicators (FSIs). While FSIs measure the level of (in)stability in a financial system, the stress uncertainty index measures the degree of predictability of (in)stability. We examine the empirical relationship between the stress uncertainty index and the real economy (industrial production, employment), and the financial markets, considering impulse response functions in a small VAR setup, and compare the results to those implied by the relationship between an existing FSI (Composite indicator of Systemic Stress by the ECB) and the real economy. In addition, we examine whether the stress uncertainty index provides value added in forecasting the real economy by looking at different forecast performance measures.

C1365: **On the use of mean square error and directional forecast accuracy for model selection: A Monte Carlo investigation**
Presenter: **Robert Kunst**, Institute for Advanced Studies, Austria
Co-authors: Mauro Costantini

A new procedure is proposed for model selection based on simultaneously targeting the mean square error and directional forecast accuracy criteria. The procedure combines the two accuracy measures using a weighting scheme for the selection of the forecasting models. Monte Carlo analysis

under different scenarios serves as a tool that assesses the strength of the procedure. To this end, we consider various time series models as generation mechanisms, in particular, time-homogeneous univariate and vector autoregressions but also generating laws that involve thresholds and structural breaks. Among the forecast models fitted to the generated data, we also consider Bayesian vector autoregressions

C0177: **The Identifying information in the forecast error variance: An application to uncertainty shocks**
Presenter: **Alessio Volpicella**, University of Surrey, United Kingdom
Co-authors: Andrea Carriero

A Structural Vector Autoregression identification scheme is developed based on inequality constraints on the Forecast Error Variance decomposition. We characterise the topological properties of this approach and provide algorithms for estimation and inference. We use this strategy to investigate the effects of uncertainty shocks on the economy by allowing for endogeneity, disentangling different sources of uncertainty, and separating uncertainty from pure financial shocks. Monte-Carlo exercises illustrate the effectiveness of this approach. Using US data, we find that some macro variables have a significant contemporaneous feedback effect on financial uncertainty, and overlooking this channel can lead to distortions in the estimated effects of uncertainty on the economy. Also, ignoring that uncertainty has heterogeneous sources biases the estimation. Finally, omitting the endogenous features of financial uncertainty leads to underestimating the effects of financial shocks on the economy. The relationship between these results and recent theoretical contributions is discussed.

CO617 Room Virtual R32 ADVANCES IN ECONOMETRICS Chair: Catherine Forbes

C1227: **Robust subset selection**
Presenter: **Ryan Thompson**, Monash University, Australia

The best subset selection (or 'best subsets') estimator is a classic tool for sparse regression, and developments in mathematical optimisation over the past decade have made it more computationally tractable than ever. Notwithstanding its desirable statistical properties, the best subsets estimator is susceptible to outliers and can break down in the presence of a single contaminated data point. To address this issue, a robust adaption of best subsets is proposed that is highly resistant to contamination in both the response and the predictors. The adapted estimator generalises the notion of subset selection to both predictors and observations, thereby achieving robustness in addition to sparsity. This procedure, referred to as 'robust subset selection' (or 'robust subsets'), is defined by a combinatorial optimisation problem for which modern discrete optimisation methods are applied. The robustness of the estimator in terms of the finite-sample breakdown point of its objective value is formally established. In support of this result, experiments on synthetic and real data are reported that demonstrate the superiority of robust subsets over best subsets in the presence of contamination. Importantly, robust subsets fare competitively across several metrics compared with popular robust adaptions of continuous shrinkage estimators.

C1272: **Efficient estimation of mixed-frequency state-space VARs: A precision-based approach**
Presenter: **Dan Zhu**, Monash University, Australia
Co-authors: Joshua Chan, Aubrey Poon

Mixed-frequency vector autoregression state-space models are now widely used for forecasting and nowcasting applications. However, despite their popularity, estimating such models can be computationally intensive. We propose a novel precision-based sampler to draw the missing observations of the low-frequency variables in these models. The newly proposed method builds on the recent advances in banded and sparse matrix algorithms for state-space models. In the simulation study, we find our new proposed method delivers superior accuracy and is computationally more efficient compared to standard filtering methods, which are commonly employed in these models. We also illustrate how our new proposed method can be applied to two popular empirical macroeconomic applications. The key insight from these two empirical applications highlights the importance of incorporating high-frequency indicators in macroeconomic models.

C1548: **A sparse dynamic factor approach to mortality modelling**
Presenter: **Jianjie Shi**, Monash University, Australia

Dynamic factor models (DFM) are an appealing and effective tool for handling a large number of time series. Despite its popularity among empirical macroeconomists, there are still some challenges that have made the DFM less favourable in practice. One challenge is the concern of over-parameterization, especially if there are many latent factors. This problem becomes even more acute when faced with high-dimensional time series. Another challenge is model interpretability which is often considered unattainable due to the lack of identifiability. Motivated by the finite mixture model, we develop a new sparse dynamic factor model (SDFM) that can achieve sparsity and enhance interpretability by classifying a set of time series into several different groups. The idea of SDFM is to simultaneously build several single-factor models while keeping the dependence structure among those factors via a dynamic mechanism. By applying the SDFM to French mortality data, we fit and forecast age-specific mortality rates parsimoniously. We compare the forecasting performance of this model against the ordinary DFM. Our results show that the sparse dynamic factor model generally provides superior forecasts when applied to French mortality data.

CC858 Room Virtual R33 CONTRIBUTIONS IN ECONOMETRIC AND FINANCIAL MODELLING Chair: Dennis Umlandt

C1223: **Dynamic mixture vector autoregressions with score-driven weights**
Presenter: **Dennis Umlandt**, University of Trier, Germany
Co-authors: Alexander Georges Gretener, Matthias Neuenkirch

A novel dynamic mixture vector autoregressive (VAR) model is proposed in which time-varying mixture weights are driven by the predictive likelihood score. Intuitively, the state weight of the k-th component VAR model in the subsequent period is increased if the current observation is more likely to be drawn from this particular state. The model is not limited to a specific distributional assumption and allows for straightforward likelihood-based estimation and inference. We conduct a Monte Carlo study and find that the score-driven mixture VAR model is able to adequately filter the mixture dynamics from a variety of different data generating processes which most other observation-driven dynamic mixture VAR models cannot cope with appropriately. Finally, we illustrate our approach by an application where we model the conditional joint distribution of economic and financial conditions.

C1448: **Unobserved heterogeneity in factor-augmented panel quantile model**
Presenter: **Wei Wang**, Shandong University, China
Co-authors: Xinbing Kong, Xiaodong Yan

Panel data models with various structural patterns, e.g., group pattern, structural break, sparsity, received increasingly more attention in statistics and econometrics. A factor-augmented panel quantile model is proposed that combines panel quantile model with factor structure that allows high-dimensional distribution-specific factors with loadings with structural breaks and sparsity. In addition, the slopes allow for unobserved grouped structures across individuals. An integrative procedure that detects the information regarding sparsity, group and structural break patterns of factor loadings and variable selection on high-dimensional covariates simultaneously via a doubly penalized hinge loss function. We use a speed iterative coordinate descent algorithm that automatically integrates structural break and group pattern factor loadings pertaining to a common one and recovers the sparsity formation of regression coefficients and loading elements. Consistency and asymptotic normality for the proposed estimators are developed. We show that the resulting estimators exhibit oracle properties, i.e., the proposed estimator is asymptotically equivalent to the oracle

estimator obtained using the known sparsity, group and structural break patterns. Furthermore, the simulation studies provide supportive evidence that the proposed method has good finite sample performance A real data empirical application has been provided to highlight the proposed method.

C1551: **Modelling dynamic multiple quantiles with exogenous factors**

Presenter: **Pierluigi Vallarino**, Aarhus BSS, Denmark

Co-authors: Alessandra Luati, Leopoldo Catania

A new dynamic model for the quantiles of the conditional distribution of a random variable is introduced. The model incorporates exogenous covariates by writing the quantiles of interest as the sum of two quantile functions: the first one subsumes past information on the variable of interest; the second function is a linear combination of baseline quantile functions, each one related to one of the predictors. The dynamics of the quantile functions rely on a set of martingale differences sequences arising from a quasi-score-driven approach. Parameters of the model are estimated through a two-stage quasi maximum likelihood estimator (2SQMLE). Consistency and asymptotic normality of the 2SQMLE are derived, and its finite sample properties are assessed through a simulation study. An empirical analysis concerning macro-financial variables shows that the default spread is relevant for predicting the left tail of the conditional distribution of the industrial production index.

CG013 Room Virtual R31 CONTRIBUTIONS IN RISK MANAGEMENT Chair: Consuelo Nava

C1259: **Risk parity strategy based on Kurtosis: Methodology and portfolio effects**

Presenter: **Consuelo Nava**, University of Turin, Italy

Co-authors: Maria Grazia Zoia, Maria Debora Braga

The distinguishing feature of portfolios based on risk-parity strategy is that of allocating wealth among asset classes in such a way that each of them contributes to the portfolio volatility to the same extent. We expand the research on risk parity with a new version of the strategy which replaces the volatility of portfolio returns with portfolio kurtosis as a reference measure. According to this approach, the investor still aims, when setting up the portfolio, at disseminating equally among asset classes the responsibility for portfolio returns' dispersion but she/he is focused on the huge dispersion as evidenced by relying on the fourth moment that puts more weight on extreme values /outcomes (either positive or negative) than standard deviation does. Closed-form expressions of the assets' contributions to portfolio kurtosis are determined. Through an application on real market data, the proposed methodology of Equally Weighted Kurtosis Contribution Portfolios is compared to the classical risk parity allocation strategy based on volatility and its peculiar properties are analyzed.

C1399: **To swing or not to swing: Reference point and professional baseball players**

Presenter: **Koji Yashiki**, Yokohama City University, Japan

Does a batting average influence the decision of whether or not to swing? We examine if the decision whether to swing is influenced by a batting average by Japanese professional baseball players using the pitch-by-pitch data. We show that when the batting average is just 0.300, the probability to swing becomes significantly lower than when it is well above or below .300. The result suggests that a .300 batting average is considered as a reference point for professional baseball players and influences their attitudes toward risk.

E1647: **Adversarial risk analysis for competitive business decisions with applications in banking**

Presenter: **Daniel Garcia Rasines**, ICMAT - CSIC, Spain

Co-authors: David Rios Insua, Roi Naveiro, Simon Rodriguez, Cesar Byron Guevara Maldonado

Most business decision-making problems take place in dynamic competitive uncertain environments. Most analyses in this area are based on game-theoretic concepts and variants, with their entailed common knowledge conditions, which are hardly tenable in the proposed domain. Novel approaches are proposed concerning adversarial risk analysis concepts, methods and tools usable in competitive business contexts. Specifically, we aim at supporting one decision-maker in undertaking its decisions in presence of other agents. Theoretically, we contribute with an alternative general approach based on adversarial risk analysis, which we consider in both static and dynamic environments. We illustrate the methods in two types of general contexts of interest in the banking sector: one where the other agents are competitors and one where the other agents are attackers.

Monday 20.12.2021 10:00 - 11:15 Parallel Session N – CFE-CMStatistics

EO068 Room Virtual R18 METHODOLOGICAL ADVANCEMENTS IN FUNCTIONAL DATA MODELS Chair: Zhenhua Lin

E0446: **A unified approach for hypothesis testing in functional linear models**
Presenter: **Zhenhua Lin**, National University of Singapore, Singapore
Co-authors: Yinan Lin

A unified approach is developed for hypothesis testing in various types of functional linear models, such as scalar-on-function, function-on-function, function-on-scalar models, that have a wide range of applications in functional data analysis. In addition, the proposed test can handle models of mixed types, such as models with both functional and scalar/vector predictors. Unlike most existing methods that rest on large-sample distributions of test statistics, the proposed method leverages the technique of bootstrapping max statistics and exploits the variance decay property that is an inherent feature of functional data to achieve superior numerical performance, especially when the sample size is limited. Theoretical guarantees on the validity and consistency of the proposed test are provided uniformly for a class of test statistics.

E0786: **Filtrated common functional principal components of multi-group functional data**
Presenter: **Shuhao Jiao**, KAUST, Saudi Arabia
Co-authors: Hernando Ombao

Local field potentials (LFPs) are signals that measure electrical activity in localized cortical regions from implanted tetrodes in the human or animal brain. These LFP signals are curves that are observed at multiple tetrodes which are spread across a patch on the surface of the cortex. Hence, they are treated as multi-group functional data. Most multi-group functional data contain both global features (which are shared in common to all curves) and isolated features (common only to a small subset of curves). One goal is to develop a procedure that captures these global features. We propose a novel tree-structured functional principal components (filt-fPC) model through low-dimensional functional representation – specifically via filtration. Ordinary fPCA can only capture major information from one population and hence fails to reveal the similarity of variation pattern across different groups. In contrast, a major advantage of the proposed filt-fPC method is the ability to extract components that are common to multiple groups and simultaneously preserves the idiosyncratic individual features of the different groups. Thus, the filt-PC method produces a parsimonious and interpretable low dimensional representation of multi-group functional data. The proposed filt-fPC method is employed to study the impact of a shock (induced stroke) on the functional organization structure of the rat brain.

E1677: **Random surface covariance estimation by shifted partial tracing**
Presenter: **Victor Panaretos**, EPFL, Switzerland

The problem of covariance estimation is considered for replicated surface-valued processes from the functional data analysis perspective. Considerations of statistical and computational efficiency often compel the use of separability of the covariance, even though the assumption may fail in practice. We consider a setting where the covariance structure may fail to be separable locally either due to noise contamination or due to the presence of a non-separable short-range dependent signal component. That is, the covariance is an additive perturbation of a separable component by a non-separable but banded component. We introduce nonparametric estimators hinging on the novel concept of shifted partial tracing, enabling computationally efficient estimation of the model under dense observation. Due to the denoising properties of shifted partial tracing, the methods are shown to yield consistent estimators even under noisy discrete observation, without the need for smoothing. Further to convergence rates and limit theorems, we show that the implementation of our estimators, including for the purpose of prediction, comes at no computational overhead relative to a separable model.

EO196 Room Virtual R20 SOME ISSUES IN BIOSTATISTICS Chair: Christiana Kartsonaki

E1350: **Estimation of incidence of early-onset invasive Group B Streptococcus disease in infants using Bayesian methods**
Presenter: **Bronner Goncalves**, London School of Hygiene & Tropical Medicine, United Kingdom

Neonatal invasive disease caused by Group B Streptococcus (GBS) leads to acute mortality and long-term morbidity. To guide the development of better prevention strategies, it is necessary to estimate the burden of this condition globally. We present a Bayesian model that estimates invasive GBS (iGBS) disease incidence in children aged 0 to 6 days. The model combines different types of epidemiological data: GBS colonization prevalence in pregnant women, risk of iGBS disease in children born to GBS-colonized mothers and direct estimates of iGBS disease incidence where available. We estimated country-specific maternal GBS colonization prevalence after adjustment for GBS detection assay. We then integrate these results with other epidemiological data and estimate the country-level incidence of iGBS disease including in countries with no studies that directly estimate incidence. Overall, we believe our method will contribute to a more comprehensive quantification of the burden of this disease, inform cost-effectiveness assessments of potential vaccines and identify areas where data are necessary.

E1359: **Predicting Helicobacter pylori serostatus: a comparison of classification algorithms**
Presenter: **Emmanuelle Dankwa**, University of Oxford, United Kingdom
Co-authors: Martyn Plummer, Christiana Kartsonaki

The Western blot test for Helicobacter pylori (H. pylori) infection, although well established, is more labour intensive and uses a larger amount of plasma than the alternative high-throughput multiplex serology test. Given that the tests differ slightly on the H. pylori proteins (antigens) considered, it was of interest to calibrate the results of multiplex serology to those of Western blot and to determine the relative importance of various antigens in determining H. pylori serostatus. We employed five classification algorithms: logistic regression (LR), random forest (RF), elastic net, Bayesian additive regressive trees (BART)and multidimensional monotone BART. These were trained on multiplex serology antigen-specific reactivity values and corresponding Western blot results. The predictive performance of models was compared using the Brier score, logarithmic score and the area under the receiver operating characteristic curve (AUC). All models showed good discriminative ability on a test set (min. and max. AUC: 95% and 97%, respectively). By the Brier score, BART displayed the best predictive performance, although the differences in performance scores of the five models were not substantial. The BART, LR and RF models showed high levels of agreement on antigen importance rankings, and results corroborated those of previous studies. This study demonstrates the utility of classification algorithms in calibrating H. pylori multiplex serology to Western blot.

E1354: **Methods for analyses of recurrent events in cohort studies**
Presenter: **Neil Wright**, University of Oxford, United Kingdom

Many epidemiological and clinical studies include analyses of events that can reoccur, such as hospitalisations, but utilise only the first occurrence for each participant and disregard subsequent events. Large cohort studies often collect data from routine sources such as electronic health records and health insurance systems, which include details of multiple events for each participant during follow-up. Non-parametric approaches to the analysis of recurrent events include estimation of incidence rates and the mean cumulative count. Poisson or negative binomial regression models can be used to compare incidence rates. There are several extensions to the Cox proportional hazards model for comparison of the hazards of recurrent events. In some studies, there are two or more types of recurrent events, or a recurrent event and a competing terminal event, and joint modelling of hazards or multi-state models are required. These various approaches and available methods for their application will be described. The challenges of applying these methods and communicating results in the context of a large cohort study will also be explored.

EO084 Room Virtual R21 STATISTICAL MODELS FOR SURVIVAL DATA II Chair: Marialuisa Restaino

E0884: **A variable selection method for high-dimensional survival data**
Presenter: **Sara Milito**, University of Salerno, Italy
Co-authors: Marialuisa Restaino, Francesco Giordano

In many studies, survival data with high-dimensional predictors are regularly collected. Models with a very large number of covariates are both infeasible to fit and likely to incur low predictability due to overfitting. The selection of significant variables plays a crucial role in estimating models and it is particularly difficult in high-dimensional settings, where the number of covariates may be greater than the sample size ($n < p$ and $n << p$). Several approaches select variables in presence of censored data are available in literature, but there is not unanimous consensus on which method outperforms the others. However, it is possible to exploit the advantages of all methods to obtain the final set of covariates as good as possible. Therefore, in order to improve the performance of variable selection methods, we propose a method that combines different procedures with subsampling, for identifying as relevant those covariates that are selected most frequently by the different variable selection methods on the subsampled data. By a simulation study, we evaluate the performance of the proposed procedure and compare it with other techniques.

E0975: **Using factor copula functions to model the association structure in right-censored survival data**
Presenter: **Roel Braekers**, Hasselt University, Belgium

In clustered right-censored survival data, both frailty and copula models are commonly used to describe the association between different lifetimes. Within the copula models, we use factor copula functions to model the structure between the different event times in a cluster. This new methodology allows for clusters to have variable sizes ranging from small to large and allows for intracluster dependence to be flexibly modeled by any parametric family of bivariate copulas. In this way, we encompass a wide range of dependence structures. The incorporation of covariates (possibly time-dependent) in the margins is also supported. We propose three estimation procedures: both a one- and two-stage parametric method and a two-stage semiparametric method where marginal survival functions are estimated by using a Cox proportional hazards model. For the parameter estimators, we prove that they are consistent and asymptotically normally distributed, and assess their finite sample behavior with simulation studies. Furthermore, we illustrate the proposed methods on a data set containing the time to the first insemination after calving in dairy cattle clustered in herds of different sizes.

E1044: **Randomizing relative treatment effects**
Presenter: **Dennis Dobler**, Vrije Universiteit Amsterdam, Netherlands

The relative treatment effect $p = P(T_{11} > T_{22}) + 0.5P(T_{11} = T_{22})$ is a single, meaningful number that quantifies whether Treatment 1 is superior to another Treatment 2 ($p > 0.5$) or not ($p <= 0.5$). The assumed data structure is that paired survival times $(T_{1i}, T_{2i}), i = 1,...,n$ are available where the first member of a pair has received Treatment 1 and the second Treatment 2. Because survival times could be independently right-censored, estimation of p is based on Kaplan-Meier estimators. Inference is achieved by means of a randomization procedure that randomly flips the treatment labels and thus artificially creates a situation in which $p = 0.5$ holds. Central limit theorems are obtained with the help of the newly developed, more general Randomization Empirical Process theory. The methodology will be illustrated with an application to data from a study on diabetic retinopathy in which the eyes of a patient underwent different treatments to prevent blindness.

EO519 Room Virtual R22 BRANCHING PROCESSES: THEORY, COMPUTATION AND APPLICATIONS II Chair: Miguel Gonzalez Velasco

E1539: **Modelling of Covid'19 pandemic using Crump-Mode-Jagers branching processes with vaccination**
Presenter: **Maroussia Slavtchova-Bojkova**, Sofia University, Bulgaria
Co-authors: Valeriya Simeonova

New results for the ongoing pandemics caused by Coronavirus SARS-CoV-2 will be presented based on the model of the general branching processes. We have developed and maintained three possible scenarios: main, optimistic and pessimistic depending on the time-varying reproduction number of the epidemic, considered as a measure of the effectiveness of the public health interventions changing rapidly to a greater or lesser degree throughout the pandemic waves. The functionality of the script allows for updating the forecasts based on the data of the ongoing spread of the epidemic. The focus is on the impact of the vaccination policies.

E1685: **Fluctuation limit theorem of critical controlled branching processes using limit theorems for martingale differences**
Presenter: **Pedro Martin-Chavez**, University of Extremadura, Spain
Co-authors: Miguel Gonzalez Velasco, Ines M del Puerto

Controlled branching processes are stochastic processes appropriate to model generation sizes in population dynamics studies where a control on the growth of population size is necessary for each generation. The main aim is to provide a Feller diffusion approximation for critical controlled branching processes. Previously, the result has been proved by using operator semigroup convergence theorems. An alternative proof is now provided making use of limit theorems for martingale differences. From a practical viewpoint, the interest in developing this result stems from its usefulness in determining the asymptotic distributions of estimators of the main parameters of a controlled branching process.

E1688: **Randomly indexed controlled branching processes**
Presenter: **Ines M del Puerto**, University of Extremadura, Spain
Co-authors: Miguel Gonzalez Velasco, Manuel Molina, George Yanev, Nikolay Yanev

A class of controlled branching processes with continuous time is introduced and some limiting distributions are obtained in the critical case. An extension of this class as regenerative controlled branching processes with continuous time is proposed and some asymptotic properties are considered.

EO252 Room Virtual R23 ECOSTA JOURNAL SESSION II Chair: Fabrizio Durante

E1742: **K-sample test of copulas**
Presenter: **Denys Pommeret**, ISFA (Lyon 1) & I2M, France

Copulas are still extensively studied and used to model the dependence of multivariate observations. Many applications can be found in the world of actuarial science by making it possible to detect mutualizable risks and not mutualizable; but also to build a well-diversified portfolio. We propose an equality test of K copulas simultaneously, when K populations are observed. We want to test the following null hypothesis $C_1 = C_2 = ... = C_K$, from K iid samples, possibly paired. We obtain the exact asymptotic null distribution of the test statistic and we prove the convergence of the test. The idea of the test is to transform the observations to uniform laws, then to use the decomposition of the density of the copula on an orthogonal Legendre polynomials basis. Returning to the copula function we obtain what is called copula coefficients which characterize each copula. The test then amounts to simultaneously comparing these coefficients. We apply this method to the "Society of Actuaries Group Medical Insurance Large Claims Database", in particular, we suggest a clustering algorithm to classify populations with similar dependence structures.

C1602: **Estimation and inference in factor copula models with exogenous covariates**
Presenter: **Alexander Mayer**, University of Cologne, Germany
Co-authors: Dominik Wied

A factor copula model is proposed in which factors are either simulable or estimable from exogenous information. Point estimation and inference

are based on the simulated method of moments approach with non-overlapping simulation draws. Consistency and limiting normality of the estimator is established and the validity of bootstrap standard errors is shown. Doing so, previous results from the literature are verified under low-level conditions imposed on the individual components of the factor structure. Monte Carlo evidence confirms the accuracy of the asymptotic theory in finite samples and an empirical application illustrates the usefulness of the model to explain the cross-sectional dependence between stock returns.

E1784: **Estimation of the complexity of a finite mixture distribution**
Presenter: **Yulia Kulagina**, ETH Zurich, Switzerland
Co-authors: Fadoua Balabdaoui, Andrei Kolar, Lilian Mueller

Mixture models occur in numerous settings including random and fixed effects models, clustering, deconvolution, empirical Bayes problems and many others. They are often used to model data originating from a heterogeneous population, consisting of several homogeneous subpopulations, and the problem of finding a good estimator for the number of components in the mixture arises naturally. Estimation of the order of a finite mixture model is a hard statistical task, and multiple techniques have been suggested for solving it. We will concentrate on several methods that have not gained much popularity but nonetheless deserve the attention of practitioners. These can be categorized into four groups: tools built upon the determinant of the Hankel matrix of moments of the mixing distribution, minimum distance estimators, likelihood ratio tests and Neural-Network-based approaches. We will address theoretical pillars underlying each of the methods and present the results of the comparative numerical study that has been conducted under various scenarios. According to the results, none of the methods proves to be a magic pill. The results uncover limitations of the techniques and provide practical hints for choosing the best-suited tool under specific conditions.

EO288 Room Virtual R24 BAYESIAN EMPIRICAL LIKELIHOOD-BASED INFERENCE METHODS Chair: Anna Simoni

E1463: **Case influence diagnostics from Bayesian empirical likelihood posteriors**
Presenter: **Catherine Forbes**, Monash University, Australia

The use of case influence diagnostics for moment condition models in a Bayesian empirical likelihood (EL) context is explored. Such models are common in Economics and related disciplines where theory implies a set of moment constraints, yet a full generative probability model is not specified. Case influence diagnostics provide a way to consider how well the moment condition model captures the variability in the data, and whether any of the observations exert a strong influence in the determination of the EL weights. The question of the influence of individual cases or of groups of cases is especially important in this context because the empirical moment conditions can be greatly affected by the presence of extreme observations. Following previous work, we develop case-influence diagnostic measures for Bayesian EL and consider appropriate low-dimensional summaries of case deletion which may be helpful in settings with many observations, parameters or moment conditions.

E1574: **Empirical likelihood for designed experiments**
Presenter: **Steven MacEachern**, The Ohio State University, United States
Co-authors: Eunseop Kim, Mario Peruggia

Experimental likelihood provides a framework that extends the use of likelihood from heavily structured parametric problems to those with minimal restrictions. The formulation of empirical likelihood allows one to focus inference on targeted quantities. We consider the application of the methods to the analysis of designed experiments. In this context, we address issues that arise due to blocking and multiple testing. Technical results identify the appropriate limiting distribution for a set of comparisons of interest. These same results suggest computational strategies that can be used for finite samples. The effectiveness of the method is demonstrated through simulation and analysis of an experiment on a commonly used pesticide. The method is shown to be robust to violations of the standard assumptions for designed experiments. The method extends to the linear mixed model.

E1684: **Bayesian estimation and comparison of conditional moment models**
Presenter: **Siddhartha Chib**, Washington University in Saint Louis, United States
Co-authors: Minchul Shin, Anna Simoni

Bayesian analysis is considered for models in which the unknown distribution of the outcomes is specified up to a set of conditional moment restrictions. The nonparametric exponentially tilted empirical likelihood function is constructed to satisfy a sequence of unconditional moments based on an increasing (in sample size) vector of approximating functions (such as tensor splines based on the splines of each conditioning variable). For any given sample size, results are robust to the number of expanded moments. We derive Bernstein-von Mises theorems for the behavior of the posterior distribution under both correct and incorrect specification of the conditional moments, subject to growth rate conditions (slower under misspecification) on the number of approximating functions. A large-sample theory for comparing different conditional moment models is also developed. The central result is that the marginal likelihood criterion selects the model that is less misspecified. We also introduce sparsity-based model search for high-dimensional conditioning variables, and provide efficient MCMC computations for high-dimensional parameters. Along with clarifying examples, the framework is illustrated with real-data applications to risk-factor determination in finance, and causal inference under conditional ignorability.

EO764 Room Virtual R25 RECENT ADVANCES IN HIGH-DIMENSIONAL STATISTICS Chair: Jun Song

E1701: **On efficient dimension reduction with respect to the interaction between two response variables**
Presenter: **Wei Luo**, Zhejiang University, China

Novel theory and methodologies for dimension reduction are introduced with respect to the interaction between two response variables, which is a new research problem that has wide applications in missing data analysis, causal inference, and graphical models, etc. We formulate the parameters of interest to be the locally and the globally efficient dimension reduction subspaces, and justify the generality of the corresponding low-dimensional assumption. We then construct estimating equations that characterize these parameters, using which we develop a generic family of consistent, model-free, and easily implementable dimension reduction methods called the dual inverse regression methods. We also build the theory regarding the existence of the globally efficient dimension reduction subspace, and provide a handy way to check this in practice. The proposal differs fundamentally from the literature of sufficient dimension reduction in terms of the research interest, the assumption adopted, the estimation methods, and the corresponding applications, and it potentially creates a new paradigm of dimension reduction research. Its usefulness is illustrated by simulation studies and a real data example at the end.

E1528: **Principal weighted least square support vector machine: An online dimension-reduction tool for binary classification**
Presenter: **Seung Jun Shin**, Korea University, Korea, South
Co-authors: Andreas Artemiou

As relevant technologies advance, steamed data that is continuously collected are frequently encountered in various applications, and the need for scalable algorithms becomes urgent. We propose the principal weighted least square support vector machine(PWLSSVM) as a novel tool for SDR in binary classification, in which most SDR methods suffer since they assume continuous Y. We further show that the PWLSSVM can be employed for the real-time SDR for the streamed data. Namely, the PWLSSVM estimator can be directly updated from the new data without having old data. We explore the asymptotic properties of the PWLSSVM estimator and demonstrate its promising performance in terms of both estimation accuracy and computational efficiency for both simulated and real data analysis.

E0979: **High-dimensional spatial quantile function-on-scalar regression**
Presenter: **Zhengwu Zhang**, UNC Chapel Hill, United States

A novel spatial quantile function-on-scalar regression model is presented, which studies the conditional spatial distribution of a high-dimensional functional response given scalar predictors. With the strength of both quantile regression and copula modeling, we are able to explicitly characterize the conditional distribution of the functional or image response on the whole spatial domain. The method provides a comprehensive understanding of the effect of scalar covariates on functional responses across different quantile levels and also gives a practical way to generate new images for given covariate values. Theoretically, we establish the minimax rates of convergence for estimating coefficient functions under both fixed and random designs. We further develop an efficient primal-dual algorithm to handle high-dimensional image data. Simulations and real data analysis are conducted to examine the finite-sample performance.

EO310 Room Virtual R26 ADVANCES IN OPTIMAL DESIGN OF EXPERIMENTS II Chair: Victor Casero-Alonso

E0759: **Precise pure-error estimation of the variance components in optimal split-plot designs**
Presenter: **Kalliopi Mylona**, King's College London, United Kingdom
Co-authors: Steven Gilmour, Peter Goos

A novel approach is presented to design split-plot experiments which ensures that the two variance components can be estimated from pure error and guarantees a precise estimation of the response surface model. Our novel approach involves a new Bayesian compound D-optimal design criterion which pays attention to both the variance components and the fixed treatment effects. One part of the compound criterion (the part concerned with the treatment effects) is based on the response surface model of interest, while the other part (which is concerned with pure-error estimates of the variance components) is based on the full treatment model. We demonstrate that our new criterion yields split-plot designs that outperform existing designs from the literature both in terms of the precision of the pure-error estimates and the precision of the estimates of the factor effects.

E0959: **Analysis of the covariance structure from the point of view of design in time-depending multiresponse models**
Presenter: **Juan M Rodriguez-Diaz**, University of Salamanca, CIF Q3718001E, Spain

The study of several characteristics of a population that depend on time will be analyzed from the point of view of design, assuming different covariance structures. The objective is to decide the most convenient moments where observation should be made in order to obtain the maximum information with limited resources (budget, time, etc.) The covariance structures will be analyzed from lowest to highest complexity, studying their influence in both the design and the estimation of the parameters of the corresponding model. Using an optimal design of experiments approach, a general method for the selection of the best temporal points will be developed. Different evolution models will be studied. Very different examples of application will be shown, from the study of two variables related to the capacity of resolution of mathematical problems, to the evolution of characteristics observed in several types of bacteria.

E1012: **Robust designs for toxicological test**
Presenter: **Sergio Pozuelo Campos**, University of Castilla-La Mancha, Spain
Co-authors: Mariano Amo-Salas, Victor Casero-Alonso

Toxicological tests are widely used to study toxicity in aquatic environments. Reproduction is a possible endpoint of this type of experiment and in this case, the response variable is counts. There exists literature about the suitable probability distribution that should be considered for analysing these data. In the theory of optimal experimental design, the assumption of this probability distribution is essential and when this assumption is not adequate, there may be a loss of efficiency in the design obtained. The main objective is to propose robust designs when there is uncertainty about the probability distribution of the response variable. The results have been applied to toxicological tests based on Ceriodaphnia Dubia and Lemna Minor, in addition to testing the properties of the designs obtained a simulation study is performed.

EO128 Room Virtual R34 OFF-THE-GRID METHODS FOR NONPARAMETRIC ESTIMATION Chair: Cristina Butucea

E0537: **SuperMix: Sparse regularization for mixtures**
Presenter: **Clement Marteau**, Universita Lyon 1, France

The statistical estimation of a discrete mixing measure μ^0 involved in a kernel mixture model is investigated. Using some recent advances in l_1-regularization over the space of measures, we introduce a data fitting and regularization convex program for estimating μ^0 in a grid-less manner from a sample of mixture law, this method is referred to as Beurling-LASSO. We derive a lower bound on the bandwidth of our data fitting term depending only on the support of μ^0 and its so-called minimum separation to ensure quantitative support localization error bounds. Under a so-called non-degenerate source condition, we derive a non-asymptotic support stability property. This latter shows that for a sufficiently large sample size n, our estimator has exactly as many weighted Dirac masses as the target μ^0, converging in amplitude and localization towards the true ones. Statistical performances of this estimator are investigated designing a so-called dual certificate, which is appropriate to our setting. The classical Gaussian distribution will be discussed.

E0923: **An off-the-grid method for the recovery of piecewise constant images in linear inverse problems**
Presenter: **Romain Petit**, INRIA Paris and Paris-Dauphine University, France
Co-authors: Vincent Duval, Yohann De Castro

In recent years, off-the-grid methods have drawn a lot of attention in the statistics and image processing community because of their improved robustness and statistical guarantees compared to their grid-based counterparts. We will describe a method to perform the recovery of piecewise constant ("cartoon") images, using the (gradient) total variation prior, in the spirit of the Rudin-Osher-Fatemi model. By exploiting the properties of the faces of the level sets of the regularizer and by relying on the Frank-Wolfe algorithm we propose a method that does not rely on a predefined grid, but adapts to the geometry of the unknown.

E1136: **Sparse dictionary learning**
Presenter: **Clement Hardy**, Ecole des Ponts Paristech/ EDF RetD, France

In many fields such as microscopy, astronomy, spectroscopy or imaging, signals that appear naturally have the structure of a sparse linear combination of parametric functions belonging to a continuous dictionary. We observe noisy measurements of such a signal discretely at given time points or continuously on a given interval and suppose that the noise level is known. The noise is supposed Gaussian and we treat simultaneously the discrete and continuous cases. We want to recover both the coefficients of the linear combination and the parameters of the functions involved. One of the first questions is whether it is possible to retrieve all the underlying parameters from the sole observation of the signal. When the parameters of the parametric functions are known the model becomes the sparse high-dimensional linear regression model. We assume that the parametric functions belong to a continuous dictionary and consider therefore a highly non-linear extension of the linear regression model. In order to retrieve all the parameters in the model, we formulate a regularized optimization problem, which despite its non-convex nature can be solved in a satisfactory way by numerical methods. Our work focuses on the behaviour of the estimators defined via the optimization problem with respect to the information we have on the signal.

EG093 Room K E. Safra (Multi-use 01) CONTRIBUTIONS IN COPULAS AND DEPENDENCE MODELLING (HYBRID) Chair: Alexandra Dias

E1515: **The accuracy of claim severity prediction using longitudinal data**
Presenter: **Alicja Wolny-Dominiak**, University of Economics in Katowice, Poland
Co-authors: Tomasz Zadlo

In the field of casualty/property insurance, an important issue is the prediction of the pure premium for the policy for the next insurance period. Given the available longitudinal claim data, the classical Buhlmann-Straub credibility model represents the next period's claim as a weighted average of historical claims arising from the experience of each risk group and the experience of the entire portfolio of policies. In this model, the fundamental assumption is claim independence. The claim dependence occurring in time is taken into account by the use of the copula. The credibility predictor is then the best predictor (conditional expectation) in the sense of the mean squared prediction. A bootstrap estimator of the predictor measure of accuracy is proposed which is based on the quantile of absolute prediction error. To obtain the value of the predictor, a Monte Carlo simulation is applied. To illustrate the bootstrap procedure in practice the portfolio of MOD risks from an insurance company operating on the Polish market is analyzed.

E1628: **Robustness methods for modeling count data with general dependence structures**
Presenter: **Marta Nai Ruscone**, Universita degli Studi di Genova, Italy
Co-authors: Dimitris Karlis

Bivariate Poisson models are appropriate for modelling paired count data. However, the bivariate Poisson model does not allow for a negative dependence structure. Therefore, it is necessary to consider alternatives. A natural way is to consider copulas to generate various bivariate discrete distributions. While such models exist in the literature, the issue of choosing a suitable copula has been overlooked so far. Different copulas lead to different structures and any copula misspecification can render the inference useless. We consider bivariate Poisson models generated with a copula and investigate its robustness under outliers contamination and model misspecification. Particular focus is on the robustness of copula related parameters. English Premier League data are used to demonstrate the effectiveness of our approach.

E1660: **On the pseudo-likelihood estimator for copula models parameters**
Presenter: **Alexandra Dias**, University of York, United Kingdom

A commonly used method for estimating dependence parameters in copula models is maximum pseudo-likelihood. It has been shown that despite its good asymptotic properties, this estimation method does not perform well when compared with methods of moments estimators for small and weakly dependent samples. We show that by changing the adjustment on the empirical distribution function the performance of the maximum pseudo-likelihood method can be improved, surpassing the performance of the methods of moments estimators. For now, the focus is on the Clayton copula model.

CO030 Room Virtual R28 GRAPHICAL MODELS AND NETWORKS ANALYSIS IN FINANCIAL APPLICATIONS Chair: Sandra Paterlini

C0207: **The impact of macroeconomic shocks on corporate investments**
Presenter: **Petre Caraiani**, Bucharest University of Economic Studies, Romania

The aim is to study how the impact of macroeconomics shocks on firm-level investments is influenced by the properties of economic networks. In this sense, we aim at using, on the one hand, data on firm-level corporate investments from the United States and, on the other hand, data on the structure of production activity as given by the BEA Input-Output tables. The first contribution is to create a new dataset that links firm-level data with data regarding the production network. The second contribution is to decompose the impact of various macroeconomic shocks on corporate investments according to their direct and network effect.

C0650: **Sparse graphical modelling via the sorted l1 - norm**
Presenter: **Sandra Paterlini**, University of Trento, Italy
Co-authors: Riccardo Riccobello, Malgorzata Bogdan, Giovanni Bonaccolto, Philipp Johannes Kremer

Sparse graphical modelling has attained widespread attention across various academic fields. We propose two new graphical model approaches, Gslope and Tslope, which provide sparse estimates of the precision matrix by penalizing its sorted L1-norm, and relying on Gaussian and t-Student data, respectively. In extensive simulation and real-world analysis, the new methods are compared to other state-of-the-art graphical modelling approaches. The results establish GSlope and TSlope as two new effective tools for sparse network estimation.

C0455: **A generalized precision matrix for t-Student distributions**
Presenter: **Karoline Bax**, University of Trento, Italy
Co-authors: Emanuele Taufer, Sandra Paterlini

For Gaussian graphical models, the precision matrix, defined as the inverse covariance matrix, is often used to express the dependence relationship between random variables. However, the Gaussian assumption is hardly satisfied in the financial context and therefore using the precision matrix might not necessarily result in a reliable and accurate picture of reality. We introduce a generalized precision matrix to overcome this issue. As fat tails are a well-known stylized fact in many financial time series, we focus on the t-Student distributions, pointing out that the behavior between random assets depends not just on the precision matrix but also on additional elements.

CO258 Room Virtual R29 SUSTAINABLE FINANCE II Chair: Monica Billio

C0576: **(In)-credibly green: Which bonds trade at a green bond premium?**
Presenter: **Carmelo Latino**, Leibniz Institute for Financial Research SAFE; Ca Foscari University of Venice, Germany
Co-authors: Julia Kapraun, Christopher Scheins, Christian Schlag

A theoretical rationale is provided for the non-existence of a Green premium by incorporating the role of trust in a model, where households have preferences for sustainable assets but do not receive any non-pecuniary utility from investing in Green bonds as soon as they do not trust the Green label. We further consider a setting where Green bonds have a real environmental impact and thereby reduce households' disutility from negative externalities. Also, the reduction in disutility happens only, if households trust in the implementation of the corresponding green project and if they can correctly assess the information on the environmental impact. To test our theoretically derived hypotheses empirically, we analyze a global sample of 1,500 Green bonds with respect to their pricing and find, on average, no significant premium on primary and secondary markets. Investors, thus, do not blindly trust all Green labels and are willing to accept lower yields only for certain types of bonds, namely those, that are perceived to be "Green-credible", either through a third-party certification of the Green label, or a listing of the bond on a dedicated Green exchange with tight listing requirements.

C0666: **A factor-based calibration model: An application to the relation between sustainability and option-implied distributions**
Presenter: **Giovanni Pianon**, Ca' Foscari University of Venice, Italy

Option-implied data represent an essential source of information for modelling the distribution of stock returns. However, as the scientific literature has shown, agents' risk preferences and irrationality can lead to a severe misalignment from the objective distribution. We propose a factor-based calibration model to correct for such a bias. By linking the calibration function parameters to a set of common and idiosyncratic factors, the model captures the time-varying nature of the forces driving the distortion of option-implied distributions. Moreover, thanks to copulas, it leverages the cross-sectional dependence of stock returns in implementing univariate calibration. A Bayesian inference approach and an efficient Monte Carlo posterior approximation allow us to deal with the model's high-dimensionality and intractability. We employed the model to perform density

forecasting and to shed light on the underlying factors determining the bias of risk-neutral densities, investigating, in particular, the role of ESG ratings and other sustainability indicators

C0744: **Inside the ESG ratings: (Dis)agreement and performance**
Presenter: **Monica Billio**, University of Venice, Italy
Co-authors: Michele Costola, Loriana Pelizzon, Carmelo Latino, Iva Hristova

The ESG rating criteria is examined used by prominent agencies and show that there is a lack of commonality in the definition of ESG (i) characteristics, (ii) attributes and (iii) standards in defining E, S and G components. We provide evidence that heterogeneity in rating criteria can lead agencies to have opposite opinions on the same evaluated companies and that agreement across those providers is substantially low. Those alternative definitions of ESG also affect sustainable investments leading to the identification of different investment universes and consequently to the creation of different benchmarks. This implies that in the asset management industry it is extremely difficult to measure the ability of a fund manager if financial performance sare strongly conditioned by the chosen ESG benchmark. Finally, we find that the disagreement in the scores provided by the rating agencies disperses the effect of preferences of ESG investors on asset prices, to the point that even when there is an agreement, it has no impact on financial performances.

CO466 Room Virtual R30 THE ECONOMETRICS OF COVID-19 PANDEMIC Chair: Sergio Scicchitano

C0204: **Parents under stress evaluating emergency childcare policies during the first COVID-19 lockdown in Germany**
Presenter: **Simone Schueller**, German Youth Institute, Germany
Co-authors: Hannah Steinberg

What are the effects of school and daycare facility closures during the COVID-19 pandemic on parental well-being and parenting behavior? Can emergency childcare policies during a pandemic mitigate increases in parental stress and negative parenting behavior? To answer these questions, the purpose is to leverage cross-state variation in emergency childcare eligibility rules during the first COVID-19 lockdown in Germany and draws on unique data from the 2019 and 2020 waves of the German AID: A family panel. Employing a DDD and IV approach, we identify medium-term ITT and LATE effects and find that while emergency care policies did not considerably affect parents life satisfaction, partnership satisfaction or mental health, they have effectively diminished harsh parenting behavior. We find partly gendered effects, specifically on paternal parenting behavior. The results suggest that decreasing parental well-being likely constitutes a general effect of the pandemic. In contrast, the observed increase in negative and potentially harmful parenting behavior is largely directly caused by school and daycare facility closures.

C0782: **Particulate matter and Covid-19 excess deaths: Decomposing long-term exposure and short-term effects**
Presenter: **Gabriele Beccari**, Tor Vergata University of Rome, Italy
Co-authors: Leonardo Becchetti, Gianluigi Conzo, Pierluigi Conzo, Davide De Santis, Francesco Salustri

The time-varying effect of particulate matter (PM) on Covid-19 deaths in Italian municipalities is investigated. We find that the lagged moving averages of PM2.5 and PM10 are significantly related to higher excess deceases during the first wave (end February-end May) of the disease, after controlling, among other factors, for time-varying mobility, regional and municipality fixed effects, the nonlinear contagion trend, and lockdown effects. The findings are confirmed after accounting for potential endogeneity, heterogeneous pandemic dynamics, and spatial correlation through pooled and fixed-effect instrumental variable estimates using municipal and provincial data. In addition, we decompose the overall PM effect and find evidence that pre-Covid long-term exposure and short-term variation during the pandemic matter, thereby supporting the two research hypotheses on the role of PM exposure. In terms of magnitude, we observe that a 1 microgram per cubic meter increase in PM2.5 leads to 20 percent more deaths in Italian municipalities, which is equivalent to a 5.9 percent increase in mortality rate.

C0943: **Labour and technology at the time of Covid-19: Can artificial intelligence mitigate the need for proximity?**
Presenter: **Sergio Scicchitano**, INAPP, Italy
Co-authors: Francesco Carbonero

Social distancing has become worldwide the key public policy to be implemented during the COVID-19 epidemic and reducing the degree of proximity among workers turned out to be an important dimension. Emerging literature looks at the role of automation in supporting the work of humans but the potential of Artificial Intelligence (AI) to influence the need for physical proximity in the workplace has been left largely unexplored. By using a unique and innovative dataset that combines data on advancements of AI at the occupational level with information on the required proximity in the job-place and administrative employer-employee data on job flows, our results show that AI and proximity stand in an inverse U-shape relationship at the sectoral level, with high advancements in AI that are negatively associated with proximity. We detect this pattern among sectors that were closed due to the lockdown measures as well as among sectors that remained open. We argue that, apart from the expected gains in productivity and competitiveness, preserving jobs and economic activities in a situation of high contagion may be the additional benefits of a policy favouring digitization.

CO162 Room Virtual R31 FINANCIAL CAPABILITY: MODELS AND EMPIRICAL EVIDENCE Chair: Sabrina Giordano

C1247: **Financial capability: A longitudinal perspective through hidden Markov models**
Presenter: **Sabrina Giordano**, University of Calabria, Italy
Co-authors: Roberto Colombi, Maria Kateri

Understanding how individuals make their economic and financial decisions, or assess their risk preferences, nowadays is one of the key priorities of organizations and governments proposing policies to reduce poverty, economic vulnerability and social exclusion. The governments are now playing an active role in meeting the financial capability challenge. In this direction, we investigate the evolution over time of the household financial capability as a latent psychological and behavioral trait that influences the household's decision-making to face financial issues. The latent financial capability is here measured in terms of two observed indicators: the self-perceived ability to make ends meet and the self-report of perceived risk related to financial investments, by choosing a category on ordinal scales. The way households disclose their perceptions can be affected by response style, an answering mechanism that induces respondents to choose middle/extreme categories of the scale or positive or negative sides of the scale regardless of the content. Our proposal is a hidden Markov model for longitudinal data with a bivariate latent Markov chain that jointly models the latent trait of interest (the financial capability) and an unobservable binary indicator of the respondent's form of answering (response style-driven or not) over time. The proposed model is fitted to ordinal longitudinal data from the Survey on Household Income and Wealth (Bank of Italy).

C1419: **Financial literacy among young people: Hints from PISA 2018.**
Presenter: **Mariangela Zenga**, Universita degli Studi di Milano-Bicocca -DISMEQ, Italy
Co-authors: Paola Bongini, Emanuela Rinaldi

The level of the financial literacy of young people in 13 OECD countries is analyzed using data from the PISA financial literacy assessment in 2018. A three-level regression model will help to explain the effects of the characteristics of the young student, of the school in which the student learns and the country in which the student lives on the performance in financial literacy.

C1634: **Does financial knowledge influence the demand for financial and insurance products in italy? A machine learning approach**
Presenter: **Susanna Levantesi**, Sapienza University of Rome, Italy

Co-authors: Giulia Zacchia

In recent years, machine learning techniques have assumed an increasingly central role in many areas of research, from computer science to medicine, including finance. We use random forest and gradient boosting techniques to investigate what influences the choice to invest the households financial wealth in insurance products, financial assets, and pension funds. Since financial knowledge is less researched as the main determinant of households insurance and financial products demand, we analyse how the level of adults' financial knowledge has a direct effect on the demand for these products using data on financial literacy and inclusion among Italian adults for 2020 and 2017. Research findings confirm that tree-based machine learning methods, such as random forest and gradient boosting techniques, can be a valuable complement to standard models (generalized linear models) and that financial knowledge makes a difference in insurance and financial products demand while controlling for other socio-economic characteristics.

CO036 Room Virtual R32 TOPICS IN THE ECONOMETRICS OF DSGE MODELS Chair: Marco Maria Sorge

C1054: **Unemployment, firm dynamics, and the business cycle**

Presenter: **Stefano Fasani**, Queen Mary University of London, United Kingdom

Co-authors: Andrea Colciago, Lorenza Rossi

A business cycle model that accounts for key business cycle properties of labor market variables and other aggregates is formulated and estimated. Three features distinguish our model (ESAM) from the standard model with SAM frictions in the labor market: frictional firm entry, endogenous product variety, and investment in two assets: stocks and physical capital. We estimate the structural parameters of the models by matching the IRFs obtained with a VAR. We identify shocks to technology, price markup, and wage bargaining power of workers by imposing sign restrictions in VAR responses. We deploy Bayesian minimum distance techniques to estimate structural parameters. ESAM accounts for the response of labor market variables such as wages, unemployment, job vacancies, and total hours, and for the response of profits and firm entry to the three shocks we identify. The success in replicating the dynamics of those variables is due to a form of endogenous wage moderation in response to technology shocks, that spreads from the extensive margin of investment. In SAM, that is the model with frictionless entry, the real wage typically displays a sharp response to shocks, that, in the case of technology shocks, leads to counterfactual responses of hours and profits. However, the improvement of ESAM over SAM is not confined to the replication of technology shocks. The statistical fit of ESAM, as measured by the marginal likelihood, is shown to be consistently higher than SAM.

C0288: **The spectral approach to linear rational expectations models**

Presenter: **Majid Al Sadoon**, Durham University, United Kingdom

Linear rational expectations models are considered in the frequency domain under general conditions. Necessary and sufficient conditions are developed for the existence and uniqueness of particular and generic systems and characterize the space of all solutions as an affine space in the frequency domain. It is demonstrated that solutions are not generally continuous with respect to the parameters of the models, invalidating mainstream frequentist and Bayesian methods. The ill-posedness of the problem motivates regularized solutions with theoretically guaranteed uniqueness, continuity, and even differentiability properties. Regularization is illustrated in an analysis of the limiting Gaussian likelihood functions of two analytically tractable models.

C1055: **Under the same (Chole)sky: DNK models, timing restrictions and recursive identification of monetary policy shocks**

Presenter: **Marco Maria Sorge**, University of Salerno, Italy

Co-authors: Giovanni Angelini

Recent structural VAR studies of the monetary transmission mechanism have voiced concerns about the use of recursive identification schemes based on short-run exclusion restrictions. The aim is to characterize the effects on impulse propagation of informational constraints embodying classical Cholesky-type timing restrictions in otherwise standard DSGE models. We formally show that timing restrictions can produce non-trivial moving average components of rational expectations solutions, or even serve as an independent source of model-based nonfundamentalness, thereby hampering impulse response analysis via VAR procedures. We then derive population conditions for the existence of VAR representations of DSGE economies exhibiting timing restrictions, and numerically explore their bearing on shock identification in a range of monetary models of the business cycle. The analysis reveals that dynamic New Keynesian models admit invertible equilibrium representations as well as fast-converging VAR coefficient matrices under empirically tenable parameterizations. This alleviates concerns about identification and lag truncation bias: low-order Cholesky-VARs do well at retrieving the true aggregate effects of monetary policy shocks in a Cholesky world.

CO390 Room Virtual R33 BAYESIAN METHODS IN FINANCIAL ECONOMETRICS: NEW DEVELOPMENTS Chair: Maria Kalli

C0507: **Efficient particle hybrid sampler for state-space models**

Presenter: **David Gunawan**, University of Wollongong, Australia

Co-authors: Robert Kohn, Christopher K Carter

Particle Markov Chain Monte Carlo (PMCMC) is a general approach to carry out Bayesian inference in non-linear and non-Gaussian state-space models. Our article shows how to scale up PMCMC in terms of the number of observations and parameters by expressing the target density of the PMCMCin terms of the basic uniform or standard normal random numbers, instead of the particles, used in the sequential Monte Carlo algorithm. Parameters that can be drawn efficiently conditional on the particles are generated by particle Gibbs. All the other parameters are drawn by conditioning on the basic uniform or standard normal random variables; e.g. parameters that are highly correlated with the states, or parameters whose generation is expensive when conditioning on the states. The performance of this hybrid sampler is investigated empirically by applying it to univariate and multivariate stochastic volatility models having both a large number of parameters and a large number of latent states and shows that it is much more efficient than competing PMCMC methods. We also show that the proposed hybrid sampler is ergodic.

C0968: **Bayesian semiparametric estimation of structural VAR models with stochastic volatility**

Presenter: **Matteo Iacopini**, Vrije Universiteit Amsterdam, Netherlands

Co-authors: Luca Rossini

The existing fully parametric Bayesian literature on structural VAR models with stochastic volatility (SVAR-SV) is extended by introducing an innovative Bayesian semiparametric framework to model high-dimensional time series of financial returns. A Bayesian nonparametric (BNP) approach based on a Dirichlet process mixture is used to flexibly model the returns distribution by also accounting for skewness and kurtosis, while the dynamics of each series volatility is modeled with a parametric structure. Our hierarchical prior overcomes overparametrization and over-fitting issues by clustering the coefficients into groups and shrinking the coefficients of each group toward a common location. An efficient Markov chain Monte Carlo sampling scheme is designed to perform inference in high-dimensional settings and provide a full characterization of parametric and distributional uncertainty. The proposed semiparametric approach is used to investigate returns predictability.

C1702: **Recurrent conditional heteroskedasticity**

Presenter: **Minh-Ngoc Tran**, University of Sydney, Australia

Co-authors: Robert Kohn, Nghia Nguyen Trong

A new class of financial volatility models, which we call the REcurrent Conditional Heteroskedastic (RECH) models, is proposed to improve both the in-sample analysis and out-of-sample forecast performance of the traditional conditional heteroskedastic models. In particular, we incorporate auxiliary deterministic processes, governed by recurrent neural networks, into the conditional variance of the traditional conditional heteroskedastic

models, e.g. the GARCH-type models, to flexibly capture the dynamics of the underlying volatility. The RECH models can detect interesting effects in financial volatility overlooked by the existing conditional heteroskedastic models such as the GARCH, GJR and EGARCH. The new models often have good out-of-sample forecasts while still explaining well the stylized facts of financial volatility by retaining the well-established structures of the econometric GARCH-type models. These properties are illustrated through simulation studies and applications to four real stock index datasets. A user-friendly software package, together with the examples, is available at github.

CC864 Room K0.19 (Hybrid 04) CONTRIBUTIONS IN APPLIED ECONOMETRICS Chair: Jose Olmo

C1542: **Utility based auction mechanism and model selection**
Presenter: **Robert Navratil**, Charles University, Faculty of Mathematics and Physics, Czech Republic
Co-authors: Jan Vecer

A novel closing auction mechanism is presented based on model prediction in the form of a distributional opinion about a random variable X. Different model opinions can be traded on a hypothetical market that trades their differences. Using a utility maximization technique, we describe such a market for any general random variable X and any utility function U. We specify the optimal behavior of agents and the total market that aggregates all available opinions and show that a correct distributional opinion realizes profit in expectation against any other opinion. Analytical solutions are available for random variables from the exponential family. We determine the distribution corresponding to the aggregated view of all available opinions. Finally, we show that the matching algorithm naturally generates a closing auction similar to NASDAQ's opening and closing cross.

C1249: **Monitoring the pandemic: A fractional filter for the COVID-19 contact rate**
Presenter: **Tobias Hartl**, University of Regensburg, Germany

The aim is to provide reliable estimates for the COVID-19 contact rate of a Susceptible-Infected-Recovered (SIR) model. From observable data on confirmed, recovered, and deceased cases, a noisy measurement for the contact rate can be constructed. To filter out measurement errors and seasonality, a novel unobserved components (UC) model is set up. It specifies the log contact rate as a latent, fractionally integrated process of unknown integration order. The fractional specification reflects key characteristics of aggregate social behavior such as strong persistence and gradual adjustments to new information. A computationally simple modification of the Kalman filter is introduced and is termed the fractional filter. It allows to estimate UC models with richer long-run dynamics, and provides a closed-form expression for the prediction error of UC models. Based on the latter, a conditional-sum-of-squares (CSS) estimator for the model parameters is set up that is shown to be consistent and asymptotically normally distributed. The resulting contact rate estimates for several countries are well in line with the chronology of the pandemic, and allow to identify different contact regimes generated by policy interventions. As the fractional filter is shown to provide precise contact rate estimates at the end of the sample, it bears great potential for monitoring the pandemic in real-time.

C1610: **Predicting housing sale prices in Germany by applying machine learning models and methods of data exploration**
Presenter: **Chong Dae Kim**, TH Koeln (Technische Hochschule Koeln), Germany
Co-authors: Nils Bedorf

The prediction of real estate prices is a popular problem in the field of machine learning and is often demonstrated. In contrast to other approaches, which regularly focus on the US market, the focus is on the biggest, German real estate dataset, with more than 1.5 million unique samples and more than 20 features. We implement and compare different machine learning models with respect to performance and interpretability to give insight into the most important properties which contribute to the sale price. The experiments suggest that the prediction of sale prices in a real-world scenario is achievable yet limited by the quality of data rather than quantity. The results show promising prediction scores but are also heavily dependent on the location, which leaves room for further evaluation.

CC865 Room Virtual R27 CONTRIBUTIONS IN MACRO AND FINANCE II (VIRTUAL) Chair: Demetris Koursaros

C0270: **Asymmetry in inflation persistence under inflation targeting**
Presenter: **Demetris Koursaros**, Cyprus University of Technology, Cyprus

The aim is to empirically document that inflation is significantly more persistent when it is below the Central Bank's target than otherwise, in five inflation targeting countries (Australia, New Zealand, Sweden, United States and the Euro-Area). We use a threshold autoregressive model to test for this asymmetry in inflation persistence; above and below some estimated threshold. We find that the threshold estimates are reasonable in light of a central bank's announced inflation target. Theoretically, we postulate that this phenomenon occurs because while forming their expectations, agents pay attention to recent observations asymmetrically along the business cycle. It is shown that a New Keynesian model with adaptive learning and an adaptive gain can explain the asymmetry in inflation persistence. Due to relatively larger forecasting errors, agents tend to put more weight on recent events in expansions, forcing inflation persistence to deteriorate. Our empirical evidence supports the theoretical findings that inflationary periods are associated with larger forecasting errors.

C1447: **Joining the gig workforce: A (potentially) one-way trip with an expensive return ticket**
Presenter: **Jessie Wang**, National Institute of Health, National Institute on Aging and RAND Corporation, United States
Co-authors: Kristina Sargent

A dynamic search and match model is proposed to explore the labor market implications of a booming gig economy. The economy has a conventional and a gig sector, with workers searching in both. There are two types of workers: those who never consider gig employment, and those who do under certain conditions. Workers match with gig positions with probability one, but face more frictions in gaining employment in the conventional sector afterwards. As a result, gig work plays an important role as an alternative to collecting unemployment insurance benefits for workers, and the gig sector absorbs labor market slack from the conventional sector. We show that the barriers to returning to the conventional market from the gig sector lead to an increased proportion of gig workers over time. By comparing the implications of the model under various levels of exposure to gig work, we explore the nature of the sector and the opportunities and consequences that come with it. The benchmark model provides insights into the rise of the gig economy, highlighting its impact on low-wage workers and the segmentation of the labor markets. Counterfactual exercises reveal the sensitivity of the sector to labor market conditions and policy interventions.

C1569: **A smooth shadow-rate dynamic Nelson-Siegel model of the yield curve at the zero lower bound**
Presenter: **Daan Opschoor**, Erasmus University Rotterdam, Netherlands
Co-authors: Michel van der Wel

A smooth shadow-rate version of the dynamic Nelson-Siegel (DNS) model is proposed to analyze the term structure of interest rates during the recent zero lower bound (ZLB) period. By relaxing the no-arbitrage restriction, our shadow-rate model becomes highly tractable with a closed-form yield curve expression and easily permits the implementation of readily available DNS extensions such as time-varying parameters, integration of macroeconomic variables and time-varying volatility. Using U.S. Treasury data, we provide clear evidence of a smooth transition of the yields entering and leaving the ZLB state. Moreover, we show that the smooth shadow-rate DNS model dominates the baseline DNS model in terms of fitting and forecasting the yield curve, while being competitive with shadow-rate affine term structure models.

CG109 Room K0.18 (Hybrid 03) CONTRIBUTIONS IN CREDIT RISK Chair: Massimiliano Caporin

C0502: **Credit rating downgrade risk on equity returns**

Presenter: **Periklis Brakatsoulas**, Charles University, Faculty of Social Sciences, Czech Republic

An asset pricing model is developed to capture credit rating downgrade risk and a new methodology is suggested to generate firm-level downgrade probabilities. Using credit transition matrices and rating histories from US issuers, we provide empirical evidence for a statistically significant positive downgrade risk premium. Stocks at a higher risk of failure tend to deliver higher returns. The performance of the model remains robust across several panel data estimation methods. Panel Granger causality test results further indicate a Granger-causal relationship from credit rating transition probabilities to excess returns. The basis for further development and empirical validation of Fama-French-type models under financial distress is provided.

C1665: **Forecasting corporate credit spreads: Regime-switching in LSTM**

Presenter: **Christina Erlwein-Sayer**, University of Applied Sciences HTW Berlin, Germany

Co-authors: Stefanie Grimm, Alexander Pieper, Rumeysa Alsac

Corporate credit spreads are modelled through a Hidden Markov model (HMM) which is based on a discretised Ornstein-Uhlenbeck model. We forecast the credit spreads within this HMM and filter out state-related information hidden in the observed spreads. We build a long short-term memory recurrent neural network (LSTM) which utilises the regime-switching information as a feature to predict the change of the credit spread. The performance of the LSTM is analysed and compared to the accuracy of an LSTM without regime-switching information. Furthermore, purely utilising the HMM forecast, the prediction of the credit spread is compared to the prediction within the LSTM. The HMM-LSTM model is calibrated on corporate credit spreads from three European countries between 2004 and 2019. Our findings show that in most cases the LSTM performance is improved when regime information is added.

C1603: **Credit risk modeling in the age of machine learning**

Presenter: **Noah Urban**, University of Duisburg-Essen, Germany

Co-authors: Martin Hibbeln, Raphael Kopp

Based on the world's largest loss database of non-retail defaults, we perform a comparative analysis of machine learning methods in credit risk modeling across the globe. We identify substantial benefits in using machine learning methods, especially tree-based methods, frequently more than doubling the performance metrics, over both simple and sophisticated benchmarks that particularly consider the specific distributions of credit risk parameters. Superior predictive abilities across many dimensions are primarily attributable to nonlinear relationships between the features and the credit risk parameters traced by methods of the explainable machine learning toolbox. Finally, we highlight important differences regarding the nature of macroeconomic features and implications of the temporal order of defaults. The results are robust to a battery of different specifications.

Monday 20.12.2021 14:45 - 16:25 Parallel Session Q – CFE-CMStatistics

EO296 Room K0.16 (Hybrid 02) ADVANCES IN BAYESIAN METHODS AND APPLICATIONS Chair: David Rossell

E0540: **Bayesian causal graphical models with purely observational data**
Presenter: **Yang Ni**, Texas A&M University, United States

A novel Bayesian causal graphical model approach is presented for reverse-engineering gene regulatory networks based on purely observational genomic data. The proposed model is provably identifiable. Empirical studies support its practical utility.

E1178: **Leveraging external data in Bayesian adaptive platform designs**
Presenter: **Alejandra Avalos Pacheco**, Harvard Medical School, Mexico
Co-authors: David Rossell, Steffen Ventz, Lorenzo Trippa

There is growing interest in trial designs that incorporate data from real-world observational studies or from previously completed trials with the goal of increasing power and reducing the sample size of clinical studies in comparison with Randomized Controlled Trials (RCT). However, if the outcome distributions of the external and internal data differ, the integration of external data may lead to biased treatment effects estimates, reduced power or increased type I error rates. We introduce a novel design that leverages external data via a Bayesian model averaging approach. The design adjusts for confounding and satisfies a set of constraints on the study's operating characteristics required by regulators. We compare two methods to perform the final analyses of the trial: i) a test based on weighted averages of p-values; ii) a non-parametric test based on permutations. We illustrate the performance of our proposed hybrid design in simulation studies based on data from real phase II and III trials.

E1240: **Bayesian learning from synthetic data**
Presenter: **Jack Jewson**, Universitat Pompeu Fabra and Barcelona Graduate School of Economics, Spain

There is significant growth and interest in the use of synthetic data as an enabler for machine learning in environments where the release of real data is restricted due to privacy or availability constraints. However, mechanisms of privacy preservation introduce artefacts in the resulting synthetic data. We use a Bayesian paradigm to characterise the updating of model parameters when learning in these settings, demonstrating that such downstream tasks can be significantly biased and that careful consideration should be given to the synthetic data generating process and learning task at hand. Recent results from general Bayesian updating allow us to propose several bias mitigation strategies inspired by decision theory, robust statistics and privatised likelihood ratios that have general applicability to differentially private synthetic data generative models. Finally, we highlight that even after bias correction significant challenges remain for the usefulness of synthetic private data generators for tasks such as prediction and inference.

E1335: **Time-varying non-linear predictions of asset returns**
Presenter: **Frank Rotiroti**, The University of Texas at Austin, United States
Co-authors: Carlos Carvalho, Jared Murray

A Bayesian approach is presented to modeling time-dependent data based on an extension of the Bayesian Additive Regression Trees (BART) model. Like BART, our approach consists of a Bayesian sum-of-trees model that constrains each tree to be a weak learner by a regularization prior; however, rather than equip each terminal node with a single mean parameter, we introduce a series of mean parameters generated according to a first-order autoregressive process. With this approach, we are better able to capture the dynamics of time-dependent data, while also taking advantage of the ability due to the BART framework to model nonlinearities and interactions among the predictors, as we demonstrate through simulation studies as well as an application to asset pricing.

EO226 Room K0.19 (Hybrid 04) STATISTICS IN NEUROSCIENCE I Chair: Jeff Goldsmith

E1093: **A sparse blind source separation method for probing human whole-brain connectomes**
Presenter: **Ying Guo**, Emory University, United States
Co-authors: Yikai Wang

In neuroscience research, imaging-based network connectivity measures have become the key for understanding brain connectomes, potentially serving as individual neural fingerprints. There are major challenges in analyzing connectivity matrices including the high dimensionality of brain networks, unknown latent sources underlying the observed connectivity, and the large number of brain connections leading to spurious findings. We propose a novel blind source separation method with low-rank structure and uniform sparsity (LOCUS) as a fully data-driven decomposition method for network measures. Compared with existing methods that vectorizes connectivity matrices ignoring brain network topology, LOCUS achieves more efficient and accurate source separation for connectivity matrices using the low-rank structure and a novel angle-based uniform sparsity regularization. We propose an efficient iterative Node-Rotation algorithm to solve the non-convex optimization problem for learning LOCUS. We illustrate LOCUS through extensive simulation studies and application to a resting-state fMRI data.

E1328: **Maximum likelihood estimation of a covariance matrix with thresholding: Application to Huntington disease**
Presenter: **Tanya Garcia**, UNC Chapel Hill, United States
Co-authors: Rakheon Kim, Mohsen Pourahmadi

The covariance matrix for a multivariate normal distribution is estimated when some entries of the matrix are zero. Compared to some existing methods such as thresholding, a positive-definite and asymptotically efficient estimator does not lose validity as a covariance matrix and provides higher confidence in estimation. However, such an estimator can be obtained only when the location of the zero entries is correctly identified. Moreover, even when the location of the zero entries is known, current approaches may fail to guarantee either positive-definiteness or asymptotic efficiency. We show that a positive-definite and asymptotically efficient estimator can always be computed by iterative conditional fitting when the location of the zero entries is known. Also, when the location of the zero entries is unknown, we propose a positive-definite thresholding estimator by combining iterative conditional fitting with thresholding and show that it is asymptotically efficient with probability tending to one. In simulation studies, the proposed estimator detected more non-zero covariances correctly, having a lower distance to the true covariance matrix than other thresholding estimators. Application to Huntington disease data detected non-zero correlations among brain regional volumes, informing which brain regions would be affected by a treatment for the disease

E1525: **Adaptive functional principal component analysis**
Presenter: **Jeff Goldsmith**, Columbia University, United States
Co-authors: Angel Garcia de la Garza, Britton Sauerbrei, Adam Hantman

Recent advances have allowed high-resolution observations of firing rates for a collection of individual neurons; these observations can provide insights into patterns of brain activation during the execution of tasks. Our data come from an experiment in which mice performed a reaching motion following an auditory cue, and contain measurements on firing rates from neuron activation in the motor cortex before and after the cue. In this setting, steep increases in firing rates after the cue are expected. Our dimension reduction technique adequately models these sharp changes over time and correctly captures these activation patterns. Initial results suggest different patterns of activation, representing the involvement of different motor cortex functions at different times in the reaching motion.

E0680: **A robust measure of effect size for neuroimage analysis**
Presenter: **Simon Vandekar**, Vanderbilt University, United States

The classical approach for testing statistical images using spatial extent inference (SEI) thresholds the statistical image based on a probability threshold (the *p*-value). This approach has an unfortunate consequence on the replicability of neuroimaging because the target set of the image is affected by the sample size – larger studies have more power to detect smaller effects. We present a general robust measure of effect size – not just applicable for neuroimaging. We use this robust effect size index with the preprocessed (ABIDE) data set, interactive visualizations, and a fully reproducible analysis pipeline to argue for thresholding statistical images by effect sizes instead of probability values. Using a constant effect size threshold means that the *p*-value threshold naturally scales with the sample size to ensure that the target set is similar across repetitions of the study that use different sample sizes. Because the statistical threshold depends on the sample size, inference procedures must be used that maintain accurate error rates at an arbitrary p-value cluster forming threshold. Future work may investigate how effect size thresholding affects SEI power in small sample sizes and meta-analytic results.

EO792 Room K0.20 (Hybrid 05) DYNAMICAL SYSTEMS IN MACHINE LEARNING Chair: Anna Korba

E0289: **Stochastic optimization with momentum: Convergence, fluctuations, and traps avoidance**
Presenter: **Anas Barakat**, Telecom Paris, IPParis, France
Co-authors: Pascal Bianchi, Walid Hachem , Sholom Schechtman

A general stochastic optimization procedure is presented by unifying several variants of the stochastic gradient descent, such as, among others, the stochastic heavy ball method, the Stochastic Nesterov Accelerated Gradient algorithm (S-NAG), and the widely used Adam algorithm. The algorithm is seen as a noisy Euler discretization of a recently introduced non-autonomous ordinary differential equation, which is analyzed in depth. Assuming that the objective function is non-convex and differentiable, the stability and the almost sure convergence of the iterates to the set of critical points are established. A noteworthy special case is the convergence proof of S-NAG in a non-convex setting. Under some assumptions, the convergence rate is provided in the form of a Central Limit Theorem. Finally, the non-convergence of the algorithm to undesired critical points, such as local maxima or saddle points, is established. Here, the main ingredient is a new avoidance of traps result for non-autonomous settings, which is of independent interest.

E0732: **Momentum residual neural networks**
Presenter: **Michael Sander**, CNRS, Projet NORIA, ENS, France
Co-authors: Pierre Ablin, Mathieu Blondel, Gabriel Peyre

The training of deep residual neural networks (ResNets) with backpropagation has a memory cost that increases linearly with respect to the depth of the network. A way to circumvent this issue is to use reversible architectures. We propose to change the forward rule of a ResNet by adding a momentum term. The resulting networks, momentum residual neural networks (Momentum ResNets), are invertible. Unlike previous invertible architectures, they can be used as a drop-in replacement for any existing ResNet block. We show that Momentum ResNets can be interpreted in the infinitesimal step size regime as second-order ordinary differential equations (ODEs) and exactly characterize how adding momentum progressively increases the representation capabilities of Momentum ResNets. Our analysis reveals that Momentum ResNets can learn any linear mapping up to a multiplicative factor, while ResNets cannot. In a learning to optimize setting, where convergence to a fixed point is required, we show theoretically and empirically that our method succeeds while existing invertible architectures fail. We show on CIFAR and ImageNet that Momentum ResNets have the same accuracy as ResNets, while having a much smaller memory footprint, and show that pre-trained Momentum ResNets are promising for fine-tuning models.

E0991: **KALE flow: A relaxed KL gradient flow for probabilities with disjoint support**
Presenter: **Pierre Glaser**, UCL, United Kingdom
Co-authors: Arthur Gretton, Michael Arbel

The gradient flow for a relaxed approximation to the Kullback-Leibler (KL) divergence between a moving source and a fixed target distribution is studied. This approximation, termed the KALE (KL approximate lower-bound estimator), solves a regularized version of the Fenchel dual problem defining the KL over a restricted class of functions. When using a Reproducing Kernel Hilbert Space (RKHS) to define the function class, we show that the KALE continuously interpolates between the KL and the Maximum Mean Discrepancy (MMD). Like the MMD and other Integral Probability Metrics, the KALE remains well defined for mutually singular distributions. Nonetheless, the KALE inherits from the limiting KL a greater sensitivity to mismatch in the support of the distributions, compared with the MMD. These two properties make the KALE gradient flow particularly well suited when the target distribution is supported on a low-dimensional manifold. Under an assumption of sufficient smoothness of the trajectories, we show the global convergence of the KALE flow. We propose a particle implementation of the flow given initial samples from the source and the target distribution, which we use to empirically confirm the KALE's properties.

E1530: **Stein optimal transport for Bayesian inference**
Presenter: **Nikolas Nuesken**, University of Potsdam, Germany

The focus is on the Stein optimal transport (Stein-OT), a novel methodology for Bayesian inference that pushes an ensemble of particles along a predefined curve of tempered probability distributions. The driving vector field is chosen from a reproducing kernel Hilbert space and can equivalently be obtained from either a suitable kernel ridge regression formulation or as an infinitesimal optimal transport map. The update equations of Stein-OT resemble those of Stein variational gradient descent (SVGD), but introduce a time-varying score function as well as specific weights attached to the particles. We will discuss the geometric underpinnings of Stein-OT and SVGD, and - time permitting - connections to MCMC and the theory of large deviations.

EO318 Room K0.50 (Hybrid 06) ADVANCES IN OPTIMAL DESIGN OF EXPERIMENTS I (VIRTUAL) Chair: Stefanie Biedermann

E0788: **Design of experiments for autoregressive networks**
Presenter: **Ben Parker**, Brunel University, United Kingdom
Co-authors: Steven Gilmour, Vasiliki Koutra

In much traditional experimental design methodology, it is assumed that experimental units are unaffected by other experimental units, or occasionally that there is some simple structure that defines their common behaviour (for example blocking). We expand recent research on designing experiments on networks. Here we develop a methodology for designing experiments on a network where experimental units are related by an autoregressive model, such that the response of each experimental unit depends on neighbouring units specified by a general adjacency matrix. For example, these may be experiments where we measure: i) the popularity of an advert that is spread on an online social network; ii) the effectiveness of an agricultural treatment where responses from one plot are correlated with their neighbours; iii) a spatial network where responses are linked based on geographical closeness. We demonstrate some simple (pseudo-Bayesian) designs on these networks. We show the importance of accounting for this autoregressive effect, and that neglecting it in the experimental design can produce very inefficient experiments.

E0878: **Partially orthogonal blocked three-level response surface designs**
Presenter: **Heiko Grossmann**, Otto-von-Guericke-University Magdeburg, Germany
Co-authors: Steven Gilmour

When fitting second-order response surface models in a hypercuboidal region of experimentation, the variance matrices of D-optimal continuous designs have a particularly attractive structure, as do many regular unblocked exact designs. Methods for constructing blocked exact designs which preserve this structure and are orthogonal, or nearly orthogonal, are developed. Partially orthogonal designs are built using a small irregular fraction of a two- or three-level design and a regular fractional factorial design as building blocks. Results are presented which relate the properties of the blocked design to these components. We show by example that partially orthogonal designs can compete with more traditional designs in terms of both efficiency and overall size of the experiment.

E0932: **Orthogonal minimally aliased response surface designs with and without two-level categorical factors**
Presenter: **Peter Goos**, KU Leuven, Belgium
Co-authors: Jose Nunez Ares, Eric Schoen

Orthogonal minimally aliased response surface or OMARS designs constitute a new family of three-level experimental designs for studying quantitative factors. Many experiments, however, also involve one or more two-level categorical factors. We derive necessary conditions for the existence of mixed-level OMARS designs and present three construction methods based on integer programming. Like the original three-level OMARS designs, the new mixed-level designs are orthogonal main-effect plans in which the main effects are also orthogonal to the second-order effects. These properties distinguish the new designs from definitive screening designs with additional two-level categorical factors and other mixed-level designs recently presented in the literature.

E0952: **Improving tests for missing-not-at-random using design of experiments**
Presenter: **Jack Noonan**, Cardiff University, United Kingdom

Missing data is known to be an inherent and pervasive problem in the process of data collection. The effects are wide-ranging and the loss of data can lead to inefficiencies and introduce bias into analyses. The specific problem of data missing not at random (MNAR) is known to be one of the most complex and challenging problems to handle in this area and testing its prevalence is of great importance. The presence of MNAR missingness can only be tested using a follow-up sample of the missing observations and therefore recovering a proportion of missing values in an efficient way could be crucial in saving the experimenters costs and time and may result in new treatments/technology reaching the public faster. We demonstrate the use of experimental design for developing a strategy to allow researchers to be in a position to be well informed about whether MNAR is a credible issue. Within a linear regression setting, we provide a proof of concept example and, using a number of newly developed approximations for the power of MNAR missingness tests, provide a number of recommendations on how the follow-up sample should be designed.

EO400 Room Virtual R18 RECENT ADVANCES IN LARGE SCALE ESTIMATION AND TESTING Chair: Trambak Banerjee

E0210: **Structural breaks in seemingly unrelated regression models**
Presenter: **Shahnaz Parsaeian**, University of Kansas, United States

An efficient Stein-like shrinkage estimator is developed for estimating the slope parameters under structural breaks in seemingly unrelated regression models, which then is used for forecasting. The proposed method is a weighted average of two estimators: a restricted estimator which estimates the parameters under the restriction of no break in the coefficients, and an unrestricted estimator which considers breakpoints and estimates the parameters using the observations within each regime. It is established that the asymptotic risk of the Stein-like shrinkage estimator is smaller than that of the unrestricted estimator which is the common method for estimating the slope coefficients under structural breaks. Furthermore, an averaging minimal mean squared error estimator is proposed where the averaging weight is derived by minimizing its asymptotic risk. The superiority of the two proposed estimators over the unrestricted estimator in terms of the mean squared forecast errors are derived. Besides, an analytical comparison between the asymptotic risks of the proposed estimators is provided. Insights from the theoretical analysis are demonstrated in Monte Carlo simulations, and on two empirical examples of forecasting U.S. industry-level inflation rates, and forecasting output growth rates of G7 countries.

E0294: **Computationally efficient penalized quantile regression**
Presenter: **Ben Sherwood**, University of Kansas, United States

Quantile regression directly models a conditional quantile. Penalized quantile regression constrains the regression coefficients similar to penalized mean regression. Quantile regression with a lasso penalty can be reframed as a quantile regression problem with augmented data and, therefore, can be formulated as a linear programming problem. If a group lasso penalty is used, then it becomes a second-order cone programming problem. These approaches become computationally burdensome for large values of n or p. Using a Huber approximation to the quantile function allows for the use of computationally efficient algorithms that require a differentiable loss function that can be implemented for both penalties. These algorithms then can be used as the backbones for implanting penalized quantile regression with other penalties such as Adaptive Lasso, SCAD, MCP and group versions of these penalties.

E0467: **Nonuniformity of p-values can occur early in diverging dimensions**
Presenter: **Emre Demirkaya**, University of Tennesse, Knoxville, United States
Co-authors: Yingying Fan, Jinchi Lv

Evaluating the joint significance of covariates is of fundamental importance in a wide range of applications. To this end, p-values are frequently employed and produced by algorithms powered by classical large-sample asymptotic theory. It is well known that the conventional p-values in the Gaussian linear model are valid even when the dimensionality is a non-vanishing fraction of the sample size. Nevertheless, they can break down when the design matrix becomes singular in higher dimensions or when the error distribution deviates from Gaussianity. A natural question is when the conventional p-values in generalized linear models become invalid in diverging dimensions. We establish that such a breakdown can occur early in nonlinear models. Simulation studies confirm the theoretical characterizations.

E1031: **Empirical Bayes control of the false discovery exceedance**
Presenter: **Pallavi Basu**, Indian School of Business, India

An empirical Bayes procedure is proposed that guarantees control of the False Discovery eXceedance (FDX) by ranking and thresholding hypotheses based on their local false discovery rate (lfdr) test statistic. In a two-group independent model or Gaussian with exchangeable hypotheses, we show that ranking by the lfdr delivers the "optimal" ranking for FDX control. We propose a computationally efficient procedure that does not empirically lose validity and power and illustrate its properties by analyzing two million stock trading strategies.

EO062 Room Virtual R20 RECENT ADVANCES IN BIOSTATISTICS Chair: Reza Drikvandi

E0259: **How to correct for baseline covariates in longitudinal clinical trials**
Presenter: **Geert Verbeke**, KU Leuven, Belgium

In clinical trials, mixed models are becoming more popular for the analysis of longitudinal data. The main motivation is often expected dropout, which can easily be handled by analysing the longitudinal trajectories. In many situations, analyses are corrected for baseline covariates such as study site or stratification variables. Key questions are then how to perform a longitudinal analysis correcting for baseline covariates, and how sensitive are the results with respect to choices made and models used. We will first present and compare a number of techniques available to correct for baseline covariates within the context of the linear mixed model for continuous outcomes. Second, we will study the sensitivity of the various techniques in case the baseline correction is based on a wrong model or does not include important covariates. Finally, our findings will be

used to formulate some general guidelines relevant in a clinical trial context. All findings and results will be illustrated extensively using data from a real clinical trial.

E0305: **Disentangling zero-inflation from overdispersion in statistical biodosimetry**
Presenter: **Jochen Einbeck**, Durham University, United Kingdom
Co-authors: Adam Errington, Jonathan Cumming, Paul Wilson
It is well known that one of the ways that overdispersion can be triggered in count data (regression) models is by the presence of 'excess' zeros, i.e. more zeros than could be expected under the employed count data distribution. This phenomenon is exploited in the detection of partial body exposures in radiation biomarkers. For instance, counts of dicentric chromosomes per cell usually adhere well to the Poisson law, unless the exposure of the body to the ionizing radiation was only partial. In that case, the non-irradiated cells will generally contribute just zero counts of chromosomal aberrations, resulting overall in an overdispersed distribution of counts. So, any overdispersion of the biomarker can be taken as evidence of partial exposure. The situation is more difficult when the radiation biomarker is overdispersed per se (that is, even under full body exposure), as, for instance, for gamma-H2AX protein foci. In this case, overdispersion is not indicative of partial exposure. Hence, what is required are methods that can separate the overdispersion resulting from the excess zeros from the 'stem' overdispersion contributed by the exposed part. We will present several approaches to this problem, among these a negative binomial version of the 'contaminated Poisson method', and a recently proposed diagnostic tool known as the 'quantile band plot'.

E0985: **Seemingly unrelated multi-state processes: A Bayesian semiparametric approach**
Presenter: **Maria De Iorio**, UCL, United Kingdom
Many applications in medical statistics as well as in other fields can be described by transitions between multiple states (e.g. from health to disease) experienced by individuals over time. In this context, multi-state models are a popular statistical technique, in particular when the exact transition times are not observed. The key quantities of interest are the transition rates, capturing the instantaneous risk of moving from one state to another. The main contribution is to propose a joint semiparametric model for several possibly related multi-state processes (Seemingly Unrelated Multi-State, SUMS, processes), assuming a Markov structure for the transitions over time. The dependence between different processes is captured by specifying a joint random effect distribution on the transition rates of each process. We assume a flexible random effect distribution, which allows for clustering of the individuals, overdispersion and outliers. Moreover, we employ a graph structure to describe the dependence among processes, exploiting tools from the Gaussian Graphical model literature. It is also possible to include covariate effects. We use our approach to model disease progression in mental health. Posterior inference is performed through a specially devised MCMC algorithm.

E1771: **Diagnostic tools for random effects in general mixed models**
Presenter: **Reza Drikvandi**, Durham University, United Kingdom
Mixed models are frequently used for the analysis of longitudinal, multilevel, clustered and other correlated data. They incorporate subject-specific random effects into the model to account for the unknown between-subject variability as well as the within-subject correlation. Since random effects are latent and unobservable variables, it is difficult to assess the random effects and their assumed distribution. There are two main challenges when working with random effects. The first challenge is to decide which random effects to include in the model. The second challenge is to check the appropriateness of the assumed distribution for random effects, which is a more difficult task. We first introduce permutation and Bayesian tests for the inclusion or exclusion of random effects from the model. We then present a likelihood-based diagnostic tool to check the adequacy of random-effects distribution. The proposed diagnostic tools can be used to assess random effects in a wide class of mixed models, including linear, generalised linear and non-linear mixed models, with univariate as well as multivariate random effects. The methods are illustrated via real data applications.

EO312 Room Virtual R21 STATISTICAL JOINT MODELING WITH LONGITUDINAL AND SURVIVAL DATA Chair: Ding Chen

E0240: **Competing risks joint model and other complex survival models using R-INLA**
Presenter: **Janet Van Niekerk**, King Abdullah University of Science and Technology, Saudi Arabia
The Integrated Nested Laplace Approximation (INLA) method is an efficient deterministic approximate Bayesian inference tool that has been used extensively in most fields of statistics and data analysis. This methodology is implemented in the R library INLA (www.r-inla.org) and is continuously reviewed and expanded. Recently, by formulating (complex) survival models as latent Gaussian models, great strides have been made in the efficient Bayesian inference of these models using INLA, instead of sampling-based approaches, which become less efficient as model complexity increases. We will present some complex survival models that we have implemented in INLA - some of which can only be fitted through MCMC besides INLA. Discrete joint models, non-linear joint models, spatial joint models, joint models with competing risks, two-part joint models, illness-death models, and spatial multi-state models are only some examples that can be efficiently estimated using the INLA methodology. The focus will be on complex joint models, even though the methodology presented can be adapted to various survival models quite trivially.

E1132: **Robust joint modelling of longitudinal data and survival data: Detection and downweighting of longitudinal measurements**
Presenter: **Freedom Gumedze**, University of Cape Town, South Africa
Mixed-effects location-scale models allow simultaneous modelling of between-subject and within-subject variability. These models include log-linear models for the between-subject and within-subject variability. The log-linear models could potentially include covariates. The models assume that the residual errors and the random effects are normally distributed. This makes them sensitive to outliers. These models have been extended to joint models of longitudinal data and time-to-event data. We explore Cook-type influence diagnostics for the mixed-effects location-scale model, assumed for the longitudinal sub-model, and an approach to down-weight outlying subjects. We illustrate the methods using data from a large cardiology clinical trial.

E1368: **Joint modeling in presence of informative censoring in palliative care studies**
Presenter: **Zhigang Li**, Department of Biostatistics, University of Florida, United States
Joint modeling of longitudinal data such as quality of life data and survival data is important for palliative care researchers to draw efficient inference because joint modeling can account for the associations between those two types of data that are commonly seen in palliative care studies. However, censoring of death times, especially informative censoring such as informative dropouts, poses challenges for modeling quality of life on a retrospective time scale. We develop a novel joint modeling approach that can address the challenge by allowing informative censoring events to be dependent on patients' quality of life through a random effect. In addition to improving the precision of estimates, our approach can provide unbiased estimates for making valid inference by appropriately modeling the informative censoring time. Model performance is assessed with a simulation study in comparison with existing approaches. A real-world study is presented to showcase the application of the new approach.

E1600: **Fitting marginalized two- part joint models to semi-continuous medical cost and survival from complex surveys**
Presenter: **Mohadeseh Shojaei Shahrokhabadi**, University of Pretoria, South Africa
Co-authors: Din Chen
To address the medical costs data problems including right skewness, clumping at zero, and censoring due to death and incomplete follow-up, Marginalized Two-part Joint Models (MTJM) have been developed. When the primary interest is to estimate covariate effects on the average costs amongst the entire population of both users and non-users, MTJM may be most useful. In the original formulation of MTJM, a Log-normal

distribution with a constant variance parameter was assumed for the positive values. We extend this model, allowing the positive values to follow a more flexible distribution- Generalized Gamma- which takes the Log-normal distribution as a special case. We use a simulation study to compare the performance of these two models with respect to bias, coverage, and efficiency. In addition, the performance of these methods is compared through application to a set of real electronic health record (EHRs) data collected in Iran. The simulation results show when the response distribution is unknown or mis-specified, that the Generalized Gamma provides a potentially more robust alternative estimator to the log-normal. For analyzing semi-continuous data with clumping at zero, researchers should consider which method is consistent with research objectives, and simultaneously appropriate for the data available.

EO416 Room Virtual R22 BAYESIAN NONPARAMETRIC METHODS IN CLASSIFICATION PROBLEMS Chair: Ramses Mena

E0401: **Bayesian scalar-on-image regression for automatically detected of regions of interest**
Presenter: **Sara Wade**, University of Edinburgh, United Kingdom

In biomedical studies, vast amounts of imaging, biological and clinical data are increasingly collected to improve understanding of diseases or conditions. In this setting, we develop scalable Bayesian scalar-on-image regression models that allow for the integration of such data. Scalar-on-image regression models utilise the entire imaging data, making it is possible to capture the complex pattern of changes associated with the disease and improve accuracy; however, the massive dimension of the images, which is often in the millions, combined with the relatively small sample size, that at best is usually in the hundreds, pose serious challenges. We propose a novel class of Bayesian nonparametric scalar-on-image regression models based on the Potts-product partition model that groups together voxels into spatially coherent clusters used as features in the regression model. This greatly eases the computational issues associated with the high-dimensional and highly-correlated inputs and allows for interpretable and reliable features that are automatically defined as the most discriminative. The posterior inference is based on a generalized Swendsen-Wang sampler, allowing efficient split-merge moves that take advantage of the spatial structure. Applications focus on early diagnosis and prognosis of Alzheimer's disease, irreversible brain disease and major international public health concerns.

E0396: **On a novel Bayesian nonparametric approach to supervised learning for binary data**
Presenter: **Jose Antonio Perusquia Cortes**, University of Kent, United Kingdom
Co-authors: Jim Griffin, Cristiano Villa

Supervised learning models provide a powerful tool for the classification of unlabeled observations. However, most of the classifiers have been built on a discriminative approach. Hence, they cannot provide an understanding of the generative process of the data. That is why the rich probabilistic background of Bayesian nonparametric models yield an interesting approach to supervised and unsupervised classification. We centre our attention on exploring a novel methodology to supervised classification for binary data using a beta compound random measure as a building block. This Bayesian nonparametric prior allows us to fully characterise the distribution of the different groups through the means of a score distribution that modifies the jumps of a directing beta process and hence, identify not only the most influential features overall but for each group as well.

E0442: **On a Dirichlet process mixture representation of phase-type distributions**
Presenter: **Luis Gutierrez**, Pontificia Universidad Catolica de Chile, Chile

An explicit representation of phase-type distributions as an infinite mixture of Erlang distributions is introduced. The representation unveils a novel and helpful connection between a class of Bayesian nonparametric mixture models and phase-type distributions. Significantly, the connection sheds some light on two hot topics, estimation techniques for phase-type distributions and the availability of closed-form expressions for some functionals related to Dirichlet process mixture models. The power of this connection is illustrated via a posterior inference algorithm to estimate phase-type distributions, avoiding some difficulties with the simulation of latent Markov jump processes, commonly encountered in phase-type Bayesian inference. On the other hand, closed-form expressions for functionals of Dirichlet process mixture models are illustrated with density and renewal function estimation.

E0927: **Bayesian modelling of sequential discoveries**
Presenter: **Tommaso Rigon**, University of Milano-Bicocca, Italy
Co-authors: David Dunson, Alessandro Zito, Otso Ovaskainen

The aim is to model the appearance of distinct tags in a sequence of labelled objects. Common examples of this type of data include words in a corpus or distinct species in a sample. These sequential discoveries are often summarised via accumulation curves, which count the number of distinct entities observed in an increasingly large set of objects. We propose a novel Bayesian nonparametric method for species sampling modelling by directly specifying the probability of a new discovery, therefore allowing for flexible specifications. The asymptotic behavior and finite sample properties of such an approach are extensively studied. Interestingly, our enlarged class of sequential processes includes highly tractable special cases. We present a subclass of models characterized by appealing theoretical and computational properties. Moreover, due to strong connections with logistic regression models, the latter subclass can naturally account for covariates. We finally test our proposal on both synthetic and real data, with special emphasis on a large fungal biodiversity study in Finland.

EO194 Room Virtual R23 ADVANCES IN BAYESIAN METHODOLOGY Chair: Victor Pena

E0246: **Bayesian fixed-domain asymptotics for covariance parameters in Gaussian process models**
Presenter: **Cheng Li**, National University of Singapore, Singapore

Gaussian process models are widely used for modeling spatial processes. We focus on the Gaussian process with isotropic Matern covariance functions. Under fixed-domain asymptotics, it is well known that when the dimension of data is less than or equal to three, the microergodic parameter can be consistently estimated with asymptotic normality while the range (or length-scale) parameter cannot. Motivated by this frequentist result, we prove a Bernstein-von Mises theorem for the covariance parameters under a Bayesian framework. Under the fixed-domain asymptotics, the posterior distribution of the microergodic parameter converges in total variation norm to a normal distribution with shrinking variance. In contrast, the posterior of the range parameter does not necessarily converge. We further propose a new property called the posterior asymptotic efficiency in linear prediction, and show that the Bayesian kriging predictor at a new spatial location with covariance parameters randomly drawn from their posterior has the same prediction mean squared error as if the true parameters were known. We illustrate these asymptotic results in numerical examples.

E0550: **Differentially private methods for managing model uncertainty in linear regression models**
Presenter: **Andres Barrientos**, Florida State University, United States
Co-authors: Victor Pena

Statistical methods for confidential data are in high demand due to an increase in computational power and changes in privacy law. Differentially private methods for handling model uncertainty in linear regression models are introduced. More precisely, we provide differentially private Bayes factors, posterior probabilities, likelihood ratio statistics, information criteria, and model-averaged estimates. Our methods are asymptotically consistent and easy to run with existing implementations of non-private methods.

E0834: **On the conjugate multivariate stochastic volatility processes**
Presenter: **Kaoru Irie**, University of Tokyo, Japan
Co-authors: Victor Pena

Some of the multivariate stochastic volatility models for a dynamic covariance matrix are known to be conjugate and appealing to the sequential analysis of steaming data. There are two established conjugate models: the inverse Wishart-matrix beta process and multiple univariate inverse gamma-beta processes combined by Bartlett decomposition. The two models are closely related but not equivalent. While they can provide identical predictive distributions, the former model has more parameters than the latter one, being conservative in retrospective uncertainty quantification. In practice, the difference between the two models cannot be assessed by their marginal likelihoods but by other model comparison measures, including the deviance information criterion and mixture estimation model. We illustrate these points by the analysis of daily returns from currency exchange rates.

E1698: **Marginalization of latent variables for correlated data**
Presenter: **Mengyang Gu**, University of California, Santa Barbara, United States
Marginalization of latent variables for correlated outcomes is discussed, including multiple time series, spatio-temporal processes, and functional data. We highlight two features of marginalization. First, we show marginalizing correlated latent variables leads to an efficient estimation of model parameters. As an example, we will introduce generalized probabilistic principal component analysis (GPPCA) to study the latent factor model for multiple correlated outcomes. The method generalizes the previous probabilistic formulation of principal component analysis (PPCA) by providing the closed-form maximum marginal likelihood estimator of the factor loadings and other parameters, where each factor is modeled by a Gaussian process. Second, we show marginalization leads to scalable computation for modeling a massive number of correlated data by Gaussian processes. Numerical studies of simulated and real data confirm the excellent finite-sample performance of the proposed approach.

EO202 Room Virtual R24 SIMULTANEOUS SUFFICIENT DIMENSION REDUCTION AND VARIABLE SELECTION Chair: Qingcong Yuan

E0315: **Fourier transform sparse inverse regression estimators for sufficient variable selection**
Presenter: **Jiaying Weng**, Bentley University, United States
Sufficient dimension reduction aims to reduce the dimension of predictors while maintaining the regression information. Recently, researchers study an impressive range of sparse inverse regression estimators. Nonetheless, conspicuously less attention has been given to the multivariate response with high-dimensional covariates settings. To fill the gap, we investigate Fourier transform inverse regression approach via regularized quadratic discrepancy functions. Theoretically, we establish the consistency and oracle property for the proposed estimators. We propose an iterated alternating direction method of multipliers (ADMM) algorithm to estimate two target parameters simultaneously. We derive the explicit solution for each step of the ADMM algorithm. Numerical studies and real data analysis confirm the theoretical properties and yield superior performance of our proposed methods. In specific, our proposal has higher support recovery rates compared to the state-of-the-art approach.

E0459: **Sufficient dimension reduction and variable selection by feature filter**
Presenter: **Pei Wang**, Miami University, United States
Sufficient dimension reduction, replacing the original predictors with a few linear combinations while keeping all the regression information, has been widely used in the past thirty years or so. We propose a new sufficient dimension reduction method, with two estimation procedures, for estimating central mean subspace through a novel approach of feature filter. The method is suitable for both univariate and multivariate responses. Asymptotic results are established. Furthermore, we provide estimation methods to determine the structural dimension, obtain a sparse estimator and deal with large p small n data. Simulations and a real data example demonstrate the efficacy of our method.

E0713: **Meta clustering for collaborative learning**
Presenter: **Chenglong Ye**, University of Kentucky, United States
Co-authors: Jie Ding, Reza Ghanadan
An emerging number of learning scenarios involve multiple learners/analysts each equipped with a unique dataset and algorithm, who may collaborate with each other to enhance their learning performance. From the perspective of a particular learner, a careless collaboration with task-irrelevant other learners is likely to incur modeling error. A crucial problem is to search for the most appropriate collaborators so that their data and modeling resources can be effectively leveraged. Motivated by this, we propose to study the problem of meta clustering, where the goal is to identify subsets of relevant learners whose collaboration will improve the performance of each individual learner. In particular, we study the scenario where each learner is performing a supervised regression, and the meta clustering aims to categorize the underlying supervised relations (between responses and predictors) instead of the raw data. We propose a general method named as Select-Exchange-Cluster (SEC) for performing such a clustering. Our method is computationally efficient as it does not require each learner to exchange their raw data. We prove that the SEC method can accurately cluster the learners into appropriate collaboration sets according to their underlying regression functions. Synthetic and real data examples show the desired performance and wide applicability of SEC to a variety of learning tasks.

E1128: **A unified framework to high dimensional sufficient dimension reduction with applications to censored data**
Presenter: **Shanshan Ding**, University of Delaware, United States
Co-authors: Wei Qian, Lan Wang
Sufficient dimension reduction (SDR) is known to be a powerful tool for achieving data reduction and visualization in regression and classification problems. We study high dimensional SDR problems and propose solutions under a unified minimum discrepancy approach with regularization. When p grows exponentially fast with n, consistency results in both central subspace estimation and variable selection are established simultaneously for important SDR methods. The proposed approach is equipped with a new algorithm to efficiently solve regularized objective functions without the need to invert a large covariance matrix. We further study a unified framework of SDR to high-dimensional survival analysis under weak modeling assumptions. This framework includes many popular survival regression models as special cases, and produces a number of practically useful outputs with theoretical guarantees, including a uniformly consistent Kaplan-Meier type estimator of the conditional distribution function of the survival time and a consistent estimator of the conditional quantile survival time in high dimension. Promising applications of our proposal are demonstrated through simulations and real data analysis on biomedical studies.

EO092 Room Virtual R25 COPULAS AND DEPENDENCE MODELLING I Chair: Piotr Jaworski

E0568: **Multivariate dependence measures revisited**
Presenter: **Martynas Manstavicius**, Vilnius University, Lithuania
Co-authors: Egle Gutauskaite
Driven by the desire to find a functional that separates two recently considered trivariate copula families, various dependence measures and several open problems are revisited. Concordance measures satisfying Taylor's axioms are not suitable for this task since in 3D they are arithmetic averages of the corresponding measures of copula 2D marginals, which in our considered case are the same for both families.

E0649: **Goodness-of-fit tests for copulae: gofCopula**
Presenter: **Martin Waltz**, Dresden University of Technology, Germany
Co-authors: Ostap Okhrin, Simon Trimborn
The last decades show an increased interest in modeling various types of data through copulae. Different copula models have been developed, which lead to the challenge of finding the best fitting model for a particular dataset. From the other side, a strand of literature developed a list of different Goodness-of-Fit (GoF) tests with different powers under different conditions. The usual practice is the selection of the best copula via the p-value

of the GoF test. Although this method is not purely correct due to the fact that non-rejection does not imply acceptance, this strategy is favored by practitioners. Unfortunately, different GoF tests often provide contradicting outputs. The proposed R-package brings under one umbrella 13 most used copulae - plus their rotated variants - together with 16 GoF tests and a hybrid one. The package offers flexible margin modeling, automatized parallelization, parameter estimation, as well as a user-friendly interface, and pleasant visualizations of the results. To illustrate the functionality of the package, two exemplary applications are provided.

E0656: **Pairwise likelihood estimation for copulas with tractable bivariate margins**
Presenter: **Jan Gorecki**, Silesian university in Opava, Czech Republic

In moderate to high dimensions, the required probability density function for the standard maximum pseudo-likelihood estimator (MPLE) of a parametric copula is often difficult to obtain, be it analytically in terms of a formula or numerically in terms of a tractable density evaluation procedure. However, the bivariate margins of such copulas are often analytically or numerically tractable. This can be exploited by the introduced pairwise pseudo-likelihood estimator (PPLE), which is studied and compared to the MPLE as an estimator for the parameters of copulas whose densities may not be numerically tractable but whose bivariate margins have tractable densities. Archimedean and related copulas serve as running examples. By simulation, the bias, root mean squared error (RMSE) and run time of the PPLE is studied for Archimedean, hierarchical Archimedean and hierarchical Archimax copulas. The PPLE is also compared to another available estimator suggested for hierarchical Archimedean copulas, the aggregated MPLE (AMPLE). The simulation results indicate that the PPLE has a comparable bias and RMSE with MPLE for those Archimedean copulas where the latter is available. For the hierarchical Archimedean and Archimax copulas where the MPLE is not easily available, the PPLE mostly outperforms the AMPLE in bias and RMSE, moreover with a clear advantage for the PPLE over the AMPLE in terms of run time.

E0762: **On copulas of a Wiener process and its running maxima and running minima processes**
Presenter: **Piotr Jaworski**, University of Warsaw, Poland

A three-variate copula of a triple of self-similar stochastic processes: a Wiener process W_t, $t \geq 0$, its running maxima process $M_t = \sup\{W_s : 0 \leq s \leq 1\}$ and its running minima process $m_t = \inf\{W_s : 0 \leq s \leq 1\}$ and its bivariate margins are the objectives. The analytical formulas for these copulas and their densities are derived, and their supports are characterized. For marginals, the Spearman ρ is calculated.

EO070 Room Virtual R26 RECENT DEVELOPMENTS IN RESPONDENT-DRIVEN SAMPLING Chair: Erica Moodie

E0458: **Learning about network features from respondent-driven sampling data**
Presenter: **Forrest Crawford**, Yale University, United States

Respondent-driven sampling (RDS) is a link-tracing procedure for surveying hidden or hard-to-reach populations in which subjects recruit other subjects via their social network. RDS recruitment does not follow a pre-specified sampling design, so making population-level inferences from samples can be difficult. We will review the graphical structure of RDS samples and present three methods for learning about network features from RDS data. First, we will show how to reconstruct the recruitment-induced subgraph of surveyed individuals probabilistically. Then, using these results, we will explain the circumstances under which it is possible to separately estimate network homophily (on measured traits in the sample) and preferential recruitment. Finally, we will discuss methods for estimating the size of the hidden population network using parametric and semi-parametric models. We apply each method to an empirical RDS dataset, including a large RDS study of injection drug users in Hartford, Connecticut, USA, in which the true population network is known. Together, these results provide a flexible set of methods that allow researchers to learn about features of a hidden social network via RDS surveys.

E0545: **Regression Modelling for Respondent-Driven Sampling**
Presenter: **Michael Rotondi**, York University, Canada
Co-authors: Lisa Avery

Respondent-driven sampling (RDS) is a relatively new technique used to recruit participants from hard to reach (hidden) populations. However, due to the statistical complexities of RDS, a number of methodological questions, including regression, remain unanswered. A simulation study was performed to evaluate the validity of various regression models that could control for the dependency between participant responses and unequal sampling probabilities in RDS. Networked populations with varying levels of homophily and prevalence, based on a known distribution of a continuous predictor were simulated and RDS samples were drawn from each population. Weighted and unweighted binomial and Poisson regression models, with and without various clustering controls were modelled for each sample to evaluate model validity. Our motivating example, examining factors associated with prevalent cardiovascular disease among the Indigenous community in Toronto is also discussed. Type-I error rates were unacceptably high for weighted regression models, dependency within the data was in general inconsequential. Even when the reported degree is accurate, as in this simulation, a low reported degree can unduly influence regression estimates. Based on the simulation results, unweighted regression should be used with RDS data and sample clustering can be ignored, at least under conditions of moderate homophily.

E0877: **Identifiability in regression methods for respondent-driven sampling**
Presenter: **Mamadou Yauck**, UQAM, Canada
Co-authors: Erica Moodie, Michael Hudgens

Respondent-driven sampling (RDS) is a form of link-tracing sampling, a technique for sampling hard-to-reach populations that aims to leverage individuals' social relationships to reach potential participants. An RDS sample represents a partially observed network of unknown dependence structures. Further, it is common to observe the social 'connectedness' of individuals with similar traits or homophily. Current analytical approaches for RDS data focus mainly on estimating means and proportions but give little technical consideration to multivariate modeling. Progress in this area is limited by a missing data problem: the observed RDS network reveals partial information about social connections between individuals in the sample. We show that the parameters of regression models are not, in general, identifiable because different full data distributions may give rise to the same observed data distribution. The lack of identification causes a violation of some model assumptions; in the modeling of the homophily effects, the conditional expectation of the error term in the linear regression, given the vector of covariates, is not zero. Thus, standard inferential methods such as maximum likelihood estimation will not in general be valid. We introduce additional assumptions to characterize the asymptotic biases of the maximum likelihood estimators of the homophily effects and the network-induced correlation parameters, and propose bias-corrected estimators.

E0990: **Clustering network tree data from respondent-driven sampling**
Presenter: **Krista Gile**, University of Massachusetts Amherst, United States

There is great interest in finding meaningful subgroups of attributed network data. There are many available methods for clustering complete networks. Unfortunately, much network data is collected through sampling, and therefore incomplete. Respondent-driven sampling (RDS) is a widely used method for sampling hard-to-reach human populations based on tracing links in the underlying unobserved social network. The resulting data, therefore, have tree structure representing a sub-sample of the network, along with many nodal attributes. We introduce an approach to adjust mixture models for general network clustering for samples collected by RDS. We apply our model to data on opioid users in New York City, and detect communities reflecting group characteristics of interest for intervention activities, including drug use patterns, social connections and other community variables.

EO396 Room Virtual R27 ADVANCES IN FUNCTIONAL DATA ANALYSIS Chair: Marzia Cremona

E0670: **Statistical shape analysis of complex networks of curves**
Presenter: **Anuj Srivastava**, Florida State University, United States

Imaging data from many applications leads to geometrical structures resembling complex pathways or curvilinear networks. We will call them "shape networks". A prominent example of a shape network is the Brain Arterial Network or BAN in the human brain, which is a complex arrangement of individual arteries, branching patterns, and inter-connectivities. Another example is a road network. Shapes or structures of these objects play an essential role in characterizing and understanding the functionality of larger systems. One would like tools for statistically analyzing shape networks, i.e., quantifying shape differences, summarizing shapes, comparing populations, and studying the effects of covariates on these shapes. The purpose is to represent and statistically analyze shape networks as "elastic shape graphs". Each elastic shape graph consists of nodes, or points in 3D, connected by some 3D curves, or edges, with arbitrary shapes. We develop a mathematical representation, a Riemannian metric, and other geometrical tools, such as computations of geodesics, means, covariances, and PCA, for helping analyze elastic shape graphs. We apply this framework to analyzing shapes of BANs taken from 92 subjects. Specifically, we generate shape summaries of BANs, perform shape PCA, and study the effects of age and gender on their shapes. We conclude that age has a clear, quantifiable effect on BAN shapes. Specifically, we find an increased variance in BAN shapes as age increases.

E0582: **False discovery rate for functional data**
Presenter: **Alessia Pini**, Universita Cattolica del Sacro Cuore, Italy
Co-authors: Niels Lundtorp Olsen, Simone Vantini

A topic that is becoming more and more popular in Functional Data Analysis is local inference, i.e., the continuous statistical testing of a null hypothesis along with a domain of interest. The principal issue in this field is the infinite amount of tests to perform, which can be seen as an extreme case of multiple comparisons problem. A number of quantities have been introduced in the literature of multivariate analysis in relation to the multiple comparisons problem. Arguably the most popular one is the False Discovery Rate (FDR), which measures the expected proportion of false discoveries among all discoveries. FDR is defined in the setting of functional data defined on a compact set of R^d, and we further generalize this definition to functional data defined on a manifold. A continuous version of the Benjamini-Hochberg method is introduced, along with a definition of adjusted p-value function. Some general conditions are stated, under which the functional Benjamini-Hochberg (fBH) procedure provides control of FDR. We show how the procedure can be plugged-in with every parametric or nonparametric pointwise test, given that such test is exact. Finally, to show the practical usefulness of our procedure, the proposed method is applied to the analysis of a data set of daily temperatures on the Earth to identify the regions where the temperature has significantly increased over the last decades.

E0764: **Data-driven identification of dynamical systems**
Presenter: **Michelle Carey**, Univerity College Dublin, Ireland
Co-authors: James Ramsay

Dynamical systems facilitate a causal explanation for the drivers and impediments of a process. But do they describe the behaviour of observed data? And how can we quantify the models' parameters that cannot be measured directly? These two questions are addressed by estimating the solution and the parameters of a linear dynamical system from incomplete and noisy observations of the processes. This methodology builds on the parameter cascading approach, where a linear combination of basis functions approximates the implicitly defined solution of the dynamical system. Then the systems' parameters are estimated so that this approximating solution adheres to the data. By taking advantage of the linearity of the system, we have simplified the parameter cascading estimation procedure, and by developing a new iterative scheme, we achieve fast and stable computation. We illustrate our approach by obtaining linear dynamical systems that represent real data from medicine, climatology and biomechanics.

E0818: **Functional data analysis characterizes the shapes of the COVID-19 epidemic in Italy**
Presenter: **Francesca Chiaromonte**, The Pennsylvania State University, United States
Co-authors: Marzia Cremona, Tobia Boschi, Jacopo DiIorio, Lorenzo Testa

COVID-19 mortality across 20 Italian regions is investigated, as well as its association with mobility, positivity, socio-demographic, infrastructural and environmental covariates. Notwithstanding limitations in accuracy and resolution of publicly available data, we pinpoint significant trends exploiting information in curves and shapes with functional data analysis. For the first epidemic wave (Feb-May 2020), we identify two starkly different patterns; an exponential one unfolding in Lombardia and the worst-hit areas of the north, and a milder, flat(tened) one in the rest of the country - including Veneto, where aggressive testing was implemented. We find that mobility and positivity predict mortality, also when controlling for relevant covariates. Among the latter, primary care appears to mitigate mortality, and contacts in hospitals, schools and workplaces aggravate it. Extending our analyses to the second epidemic wave (Oct 2020-Feb 2021) we find differences in mobility restrictions compared to the first, but we confirm a strong role for mobility and a marked heterogeneity in mortality patterns across the country. FDA techniques could capture additional signals if applied to richer data.

EO402 Room Virtual R28 BAYESIAN METHODS IN STRUCTURED DATA AND HIGH-DIMENSIONAL PROBLEMS Chair: Nilabja Guha

E0562: **Nonparametric group variable selection with multivariate response for connectome-based prediction of cognitive scores**
Presenter: **Arkaprava Roy**, University of Florida, United States

Possible relations between the structural connectome and cognitive profiles are studied using a multi-response nonparametric regression model under group sparsity. The aim is to identify the brain regions having a significant effect on cognitive functioning. The cognitive profiles are measured in terms of seven cognitive test scores from NIH toolbox of cognitive battery. The structural connectomes are represented by adjacency matrices. Most existing works consider the upper or lower triangular section of these adjacency matrices as predictors. An alternative characterization of the connectivity properties is available in terms of the nodal attributes. We consider nine different attributes for each brain region as our predictors. These nodal graph metrics may naturally be grouped together for each node, motivating us to introduce group sparsity for feature selection. We propose Russian RBF-nets with a novel group sparsity inducing prior to model the unknown mean functions. The covariance structure of the multivariate response is characterized in terms of a linear factor modeling framework. Applying our proposed method to a Human Connectome Project (HCP) dataset, we identify the important brain regions and nodal attributes for cognitive functioning, as well as identify interesting low-dimensional dependency structures among the cognition related test scores.

E1304: **Shrinkage on simplex: Bayesian inference for sparse and structured compositional data**
Presenter: **Jyotishka Datta**, Virginia Polytechnic Institute and State University, United States

Sparse signal recovery remains an important challenge in large scale data analysis and global-local (G-L) shrinkage priors have undergone an explosive development in the last decade in both theory and methodology. These developments have established the G-L priors as the state-of-the-art Bayesian tool for sparse signal recovery as well as default non-linear problems. While there is a huge literature proposing elaborate shrinkage and sparsity priors for high-dimensional real-valued parameters, there has been limited consideration of discrete data structures. We will survey the recent advances in G-L shrinkage priors, focusing on the optimality of these priors for both continuous as well as quasi-sparse count data. We will discuss an extension to discrete data structures including sparse compositional data, routinely observed in microbiomics. We will discuss the methodological challenges with the Dirichlet distribution as a shrinkage prior for high-dimensional probabilities for its inability to adapt to an

arbitrary level of sparsity, and propose to address this gap by using a new prior distribution, specially designed to enable scaling to data with many categories. We will provide some theoretical support for the proposed methods and show improved performance in several simulation settings and application to microbiome data

E1378: **Bayesian semiparametric longitudinal functional mixed models with locally informative predictors**
Presenter: **Abhra Sarkar**, The University of Texas at Austin, United States

A flexible Bayesian semiparametric mixed model is presented for longitudinal functional data analysis in the presence of potentially high-dimensional categorical covariates. The proposed method allows the fixed effects components to vary between dependent random partitions of the covariate space at different time points. The mechanism not only allows different sets of covariates to be included in the model at different time points but also allows the selected predictors' influences to vary flexibly over time. Smooth time-varying additive random effects are used to capture subject-specific heterogeneity. We establish posterior convergence guarantees for both function estimation and variable selection. We design a Markov chain Monte Carlo algorithm for posterior computation. We evaluate the methods empirical performances through synthetic experiments and demonstrate its practical utility through real-world applications.

E1386: **An approximate Bayesian approach to covariate dependent graphical modeling**
Presenter: **Debdeep Pati**, Texas A&M University, United States
Co-authors: Sutanoy Dasgupta, Bani Mallick, Prasenjit Ghosh

Gaussian graphical models typically assume a homogeneous structure across all subjects, which is often restrictive in applications. We propose a weighted pseudo-likelihood approach for graphical modeling which allows different subjects to have different graphical structures depending on extraneous covariates. The pseudo-likelihood approach replaces the joint distribution by a product of the conditional distributions of each variable. We cast the conditional distribution as a heteroscedastic regression problem, with covariate-dependent variance terms, to enable information borrowing directly from the data instead of a hierarchical framework. This allows independent graphical modelling for each subject, while retaining the benefits of a hierarchical Bayes model and being computationally tractable. An efficient embarrassingly parallel variational algorithm is developed to approximate the posterior and obtain estimates of the graphs. Using a fractional variational framework, we derive asymptotic risk bounds for the estimate in terms of a novel variant of the alpha-Renyi divergence. We theoretically demonstrate the advantages of information borrowing across covariates over independent modelling across covariates. We show the practical advantages of the approach through simulation studies and illustrate the dependence structure in protein expression levels on breast cancer patients using CNV information as covariates.

EO487 Room Virtual R29 STATISTICAL METHODS FOR HIGH-DIMENSIONAL AND DEPENDENT DATA Chair: Yumou Qiu

E1266: **Inference on multi-level partial correlations based on multi-subject time series data**
Presenter: **Yumou Qiu**, Iowa State University, United States

Partial correlations are commonly used to analyze the conditional dependence among variables. We propose a hierarchical model to study both the subject and population-level partial correlations based on multi-subject time series data. Multiple testing procedures adaptive to temporally dependent data with false discovery proportion control are proposed to identify the nonzero partial correlations in both the subject and population levels. A computationally feasible algorithm is developed. Theoretical results and simulation studies demonstrate the good properties of the proposed procedures. We illustrate the application of the proposed methods in a real example of brain connectivity on fMRI data from normal healthy persons and patients with Parkinson's disease.

E1334: **Two-sample mean test for high-dimensional time series**
Presenter: **Shuyi Zhang**, East China Normal University, China
Co-authors: Song Xi Chen, Yumou Qiu

Two population means are tested with high-dimensional temporally dependent data. To eliminate the bias caused by the temporal dependence among the time-series observations, a band-excluded U-statistic (BEU) is proposed to estimate the squared Euclidean distance between the two means, which excludes cross-products of data vectors between temporally close time points. The asymptotic normality of the BEU statistic is derived under the high dimensional setting with "spatial" (column-wise) and temporal dependence. An estimator built on the kernel smoothing over cross-time covariances is developed to estimate the variance of the BEU-statistic, which facilitates a test procedure based on the BEU statistic. The proposed test is nonparametric and adaptive to a wide range of dependence and dimensionality, and has attractive power properties relative to a self-normalized test. Numerical analysis and a real data analysis on a meteorological data-set were conducted to demonstrate the performance and utility of the proposed test.

E1374: **Forward selection in ultra-high dimensional functional concurrent models with applications to functional GWAS**
Presenter: **Lily Wang**, George Mason University, United States
Co-authors: Shan Yu, Rodrigo Plazola-Ortiz, Yehua Li

In a functional genome-wide association study (fGWAS) dataset from the Alzheimer's Disease Neuroimage Initiative (ADNI), the longitudinal Alzheimer Disease (AD) biomarkers are modeled by a class of concurrent functional linear models. The model includes the functional effects of environmental covariates, ultra-high dimensional genetic covariates and their interactions. We approximate the coefficient functions using B-splines, and select important main effects and interactions using a forward selection procedure based on a functional Bayesian Information Criterion (fBIC). The proposed fBIC is adaptive to both sparse and dense functional data, leads to a consistent variable selection procedure when the dimension of covariates is fixed; enjoys the sure screening property and leads to less false positive than existing methods when the covariate dimensionality is ultra-high. Simulation studies confirm the properties and the analysis of the ADNI data leads to new findings on AD-related environmental, genetic and interaction effects.

E1440: **A new smoothing method for 3D imaging data: Efficiency vs. accuracy**
Presenter: **Xinyi Li**, Clemson University, United States
Co-authors: Shan Yu, Yueying Wang, Guannan Wang, Lily Wang

Over the past two decades, increased demand for 3D visualization and simulation software is seen in medicine, architectural design, engineering, and many other areas, which have boosted the investigation of geometric data analysis and raised the demand for further advancement in statistical analytic approaches. We propose a class of spline smoothers appropriate for approximating geometric data over 3D complex domains, which can be represented in terms of a linear combination of spline basis functions with some smoothness constraints. We start with introducing the tetrahedral partitions, Barycentric coordinates, Bernstein basis polynomials, and trivariate spline on tetrahedra. Then, we propose a penalized spline smoothing method for identifying the underlying signal in a complex 3D domain from potential noisy observations. Simulation studies are conducted to compare the proposed method with traditional smoothing methods on 3D complex domains.

EO527 Room Virtual R30 CURRENT DEVELOPMENTS IN IMAGING DATA ANALYSIS Chair: Zhengwu Zhang

E0307: **Genetic underpinnings of brain structural connectome for young adults**
Presenter: **Yize Zhao**, Yale University, United States

With distinct advantages in power over behavioral phenotype, brain imaging traits have become emerging endophenotypes to dissect molecular contribution to behaviors and neuropsychiatric illness. Among different imaging features, brain structural connectivity which summarizes whole-

brain anatomical neural connections is one of the most cutting edges while under-investigated traits; and the genetic influence on the shifts of structural connectivity remains highly elusive. Relying on a landmark imaging genetics study for young adults, we develop a biologically plausible brain network response shrinkage model to comprehensively characterize the relationship between high dimensional genetic variants and the structural connectome phenotype. Under a unified Bayesian framework, we accommodate the topology of brain network and biological architecture within the genome; and eventually, establish a mechanistic mapping between genetic biomarkers and the associated brain sub-network units. An efficient expectation-maximization algorithm is developed to estimate the model parameters and ensure computing feasibility. We show the superiority of our method in extensive simulations. In the application to the Human Connectome Project Young Adult (HCP-YA) data, we establish the genetic underpinnings which are highly interpretable under functional annotation and brain tissue eQTL analysis, for the brain white matter tract sub-networks concentrating on the hippocampus and between hemispheres.

E1183: **Optimal generalized tensor estimation with applications to 4D-STEM image denoising**
Presenter: **Anru Zhang**, Duke University, United States

A general framework is introduced for generalized low-rank tensor learning problems, which includes many important instances arising from applications in computational imaging, genomics, network analysis, etc. To overcome the difficulty of non-convexity in these problems, we introduce a unified tensor approach that adapts to the underlying low-rank structure. Under mild conditions of rank-restricted convexity and smoothness on the loss function, we establish the upper bound on the statistical error and the linear rate of computational convergence through a general deterministic analysis. Then we further consider a suite of generalized tensor learning problems, including Gaussian tensor denoising, Poisson, binomial tensor PCA, and linear regression. Next, we apply the proposed approach to 4D-Scanning Transmission Electron Microscopy (4D-STEM) imaging analysis. For the 4D-STEM imaging data, due to the substantial noise brought up by photon-limited imaging technique, adequate and sufficient denoising often becomes the crucial first step in the analysis. Through the proposed tensor-based methods, we are able to achieve significantly better denoising performance and obtain smaller image reconstruction errors compared to the classic matrix-based denoising methods.

E1209: **Adaptive frequency band analysis for functional time series**
Presenter: **Scott Bruce**, Texas A&M University, United States
Co-authors: Pramita Bagchi

The frequency-domain properties of nonstationary functional time series often contain valuable information. These properties are characterized through their time-varying power spectrum. Practitioners seeking low-dimensional summary measures of the power spectrum often partition frequencies into bands and create collapsed measures of power within bands. However, standard frequency bands have largely been developed through manual inspection of time series data and may not adequately summarize power spectra. We propose a framework for adaptive frequency band estimation of nonstationary functional time series that optimally summarizes the time-varying dynamics of the series. We develop a scan statistic and search algorithm to detect changes in the frequency domain. We establish the theoretical properties of this framework and develop a computationally-efficient implementation. The validity of our method is also justified through numerous simulation studies and an application to analyzing electroencephalogram data in participants alternating between eyes open and eyes closed conditions.

E1442: **Big imaging data learning: A parallel solution**
Presenter: **Shan Yu**, University of Virginia, United States
Co-authors: Guannan Wang, Lei Gao, Lily Wang

Nowadays, we are living in the era of "Big Data." A significant portion of big data is big imaging data captured through advanced technologies. Explosive growth in imaging data emphasizes the need for developing new and computationally efficient methods and credible theoretical support tailored for analyzing such large-scale data. Parallel statistical computing has proved to be a handy tool when dealing with big data. In general, it uses multiple processing elements simultaneously to solve a problem. However, it is hard to execute the conventional spline regressions in parallel. We develop a novel parallel smoothing technique for generalized partially linear spatially varying coefficient models, which can be used under different hardware parallelism levels. Moreover, conflated with concurrent computing, the proposed method can be easily extended to the distributed system. Regarding the theoretical support of estimators from the proposed parallel algorithm, we first establish the asymptotical normality of linear estimators. Secondly, we show that the spline estimators reach the same convergence rate as the global spline estimators. The newly developed method is evaluated through several simulation studies and an analysis of the ADNI data.

EO134 Room Virtual R31 HIGH-DIMENSIONAL INFERENCE IN GENERALIZED LINEAR MODELS Chair: Fadoua Mohr

E1144: **Post-selection inference on high-dimensional varying-coefficient generalized linear models**
Presenter: **Ran Dai**, University of Nebraska Medical Center, United States
Co-authors: Cheng Zheng

Generalized linear models (GLMs) are important parametric extensions of linear models. Varying-coefficient modeling is frequently used in capturing the dynamics of the impact of the covariates. We study high-dimensional varying-coefficient generalized linear models, which allow us to capture non-stationary effects of the input variables across time. We develop new tools for the statistical inference that allow us to construct valid confidence intervals and honest tests for nonparametric coefficients at fixed varying-coefficient indices. The focus is on inference in a high-dimensional setting where the number of input variables exceeds the sample size. Performing statistical inference in this regime is challenging due to the usage of model selection techniques in estimation. Nevertheless, we develop valid inferential tools that are applicable to a wide range of data generating processes and do not suffer from biases introduced by model selection. We performed numerical simulations to demonstrate the finite sample performance of our method and we also illustrate the application with a real data example.

E1452: **Inference for the case probability in high-dimensional logistic regression**
Presenter: **Zijian Guo**, Rutgers University, United States

Labeling patients in electronic health records with respect to their statuses of having a disease or condition, i.e. case or control statuses, has increasingly relied on prediction models using high-dimensional variables derived from structured and unstructured electronic health record data. A major hurdle currently is a lack of valid statistical inference methods for the case probability. Considering high-dimensional sparse logistic regression models for prediction, we propose a novel bias-corrected estimator for the case probability through the development of linearization and variance enhancement techniques. We establish the asymptotic normality of the proposed estimator for any loading vector in high dimensions. We construct a confidence interval for the case probability and propose a hypothesis testing procedure for patient case-control labelling. We demonstrate the proposed method via extensive simulation studies and application to real-world electronic health record data.

E1736: **Optimal ranking in crowdsourcing**
Presenter: **Alexandra Carpentier**, Universitaet Potsdam, Germany
Co-authors: Emmanuel Pilliat, Nicolas Verzelen

Consider a crowdsourcing problem where we have n experts and d tasks. The average ability of each expert for each task is stored in an unknown matrix M, which is only observed in noise and incompletely. We make no (semi) parametric assumptions, but assume that both experts and tasks can be perfectly ranked: so that if an expert is better than another, she performs on average better on all tasks than the other - and that the same holds for the tasks. This implies that if the matrix M is permuted so that the experts and tasks are perfectly ranked, then the permuted matrix M is bi-isotonic. We focus on the problem of recovering the optimal ranking of the experts in the l_2 norm, when the questions are perfectly ranked.

We provide a minimax-optimal and computationally feasible method for this problem, based on hierarchical clustering, PCA, and the exchange of information among the clusters. We prove, in particular, - in the case where d is larger than n- that the problem of estimating the expert ranking is significantly easier than the problem of estimating the matrix M.

E1739: **Inference in high-dimensional single-index models under symmetric designs**
Presenter: **Moulinath Banerjee**, University of Michigan, United States

The problem of statistical inference for regression coefficients in a high-dimensional single-index model is considered. Under elliptical symmetry, the single index model can be reformulated as a proxy linear model whose regression parameter is identifiable. We construct estimates of the regression coefficients of interest that are similar to the de-biased lasso estimates in the standard linear model and exhibit similar properties: $\sqrt{n}$-consistency and asymptotic normality. The procedure completely bypasses the estimation of the unknown link function, which can be extremely challenging depending on the underlying structure of the problem. Furthermore, under Gaussianity, we propose more efficient estimates of the coefficients by expanding the link function in the Hermite polynomial basis. Finally, we illustrate our approach via carefully designed simulation experiments.

EO657 Room Virtual R34 ADVANCES IN THE ANALYSIS OF QUANTILES, EXPECTILES AND EXTREMILES Chair: Abdelaati Daouia

E0977: **Extreme expectile regression in heavy-tailed regression models**
Presenter: **Yasser Abbas**, Fondation Jean-Jacques Laffont, France

Studying rare events at the tails of heavy-tailed distributions is a burgeoning science and has many applications both in and out of finance. Most attempts to tackle the subject involve quantile regression, which usually offers a natural way of examining the impact of covariates at different levels of the dependent variable. We argue, however, that quantiles are not well equipped to deal with sparsity around the tails, especially in the active field of risk management, and motivate their least-square analogues, expectiles, as a more appropriate alternative. We introduce versatile estimators of tail conditional expectiles under an extremal additive regression model with heavy-tailed regression noise and derive their asymptotic properties in a general setting. We then tailor the discussion to the local linear estimation approach. We showcase the performance of our procedures in a detailed simulation study and apply them to a concrete dataset.

E0931: **Parametric measures of variability induced by risk measures**
Presenter: **Fabio Bellini**, University of Milano-Bicocca, Italy
Co-authors: Tolulope Fadina, Ruodu Wang, Yunran Wei

A general framework is presented for a comparative theory of variability measures, with a particular focus on the recently introduced one-parameter families of inter-Expected Shortfall differences and inter-expectile differences, which are explored in detail and compared with the widely known and applied inter-quantile differences. From the mathematical point of view, our main result is a characterization of symmetric and comonotonic variability measures as mixtures of inter-Expected Shortfall differences, under a few additional technical conditions. Further, we study the stochastic orders induced by the pointwise comparison of inter-Expected Shortfall and inter-expectile differences, and discuss their relationship with the dilation order. From the statistical point of view, we establish asymptotic consistency and normality of the natural estimators and provide a rule of the thumb for cross-comparisons. Finally, we study the empirical behaviour of the considered classes of variability measures on the SP500 Index under various economic regimes, and explore the comparability of different time series according to the introduced stochastic orders.

E0765: **Estimation of the largest tail-index and extreme quantiles from a mixture of heavy-tailed distributions**
Presenter: **Stephane Girard**, Inria, France
Co-authors: Emmanuel Gobet

The estimation of extreme quantiles requires adapted methods to extrapolate beyond the largest observation of the sample. Extreme-value theory provides a mathematical framework to tackle this problem together with statistical procedures based on the estimation of the so-called tail-index describing the distribution tail. We focus on heavy-tailed distributions and consider the case where the observations at hand are related to statistical models with different tail-index, a.k.a. as a mixture of heavy-tail models, and for conservative risk management reasons, we are interested in the largest tail-index. In such a mixture situation, usual extreme-value estimators suffer from a strong bias, which may induce in turn a strong bias when quantifying tail risk in this mixture model. We propose several methods to mitigate this bias under mild assumptions on the mixture distribution. Their asymptotic properties are established and their finite sample performance is illustrated both on simulated and real financial data.

E0886: **Extremile regression**
Presenter: **Gilles Stupfler**, ENSAI - CREST, France
Co-authors: Abdelaati Daouia, Irene Gijbels

Regression extremiles define a least-squares analogue of regression quantiles. They are determined by weighted expectations rather than tail probabilities, and enjoy various closed-form expressions and interpretations. Of special interest is their intuitive meaning in terms of expected minima and expected maxima. Their use appears naturally in any decision theory where the severity of tail observations, rather than their relative frequency, is of utmost interest. In risk management, for instance, quantiles rely only on the probability of tail losses and not on their values. They also fail to fulfil the coherency axiom. Extremiles are perfectly reasonable alternatives in both of these respects. We provide the first detailed study exploring implications of the extremile terminology in a general setting of presence of covariates. We follow two paths for estimating conditional extremiles and deriving the asymptotic normality of their estimators. One is based on their characterization as the weighted average of all regression quantiles, and the other relies on local linear (least squares) check function minimization. We also extend extremile regression far into the tails of heavy-tailed distributions. Extrapolated estimators are constructed and their asymptotic extreme value theory is developed. Some applications to real data are provided.

EO834 Room Virtual R36 NEW CHALLENGES ON CHANGE-POINT DETECTION (VIRTUAL) Chair: Andreas Anastasiou

E0468: **Change point ideas in multiple testing: Estimating the proportion of false null hypotheses**
Presenter: **Anica Kostic**, London School of Economics and Political Science, United Kingdom
Co-authors: Piotr Fryzlewicz

Estimating the proportion of false null hypotheses among a large number of independently tested hypotheses is an important problem in multiple testing literature. In the sequence of sorted p-values (p-value plot), false null p-values tend to be smaller and concentrated at the beginning. This suggests an approximate piecewise linear shape of the p-value plot, with a change-point in slope separating small from large p-values. We propose a method for estimating the false null proportion (Difference Of Slopes) that utilises the idea of estimating that change-point in the p-value plot.

E0606: **A changepoint approach to modelling soil moisture dynamics**
Presenter: **Mengyi Gong**, Lancaster University, United Kingdom
Co-authors: Rebecca Killick, Christopher Nemeth

Soil moisture is an important measure of soil health that scientists model via soil dry-down curves. The typical modelling process requires manually separating the soil moisture time series into segments representing the drying process and fitting exponential decay models to them. This can be time-consuming for a large data set. The result is a static overview of the dry-down property. Motivated by the spike-train problem in neuroscience, we propose a novel changepoint-based approach to automatically identify structural changes in the soil drying process. Changes caused by sudden

rises in soil moisture over a long time series are captured and the parameters characterising the drying processes are estimated simultaneously. We allow segment-specific parameters to capture potential temporal variations in the drying process. The method can be considered as a complement to the conventional soil dry-down modelling. An algorithm based on the penalised exact linear time (PELT) method was developed to identify the changepoints. A simulation study was carried out to access the performance of the method. The result demonstrated its ability to locate structure changes and retrieve key parameters. The method was applied to the 2-year hourly soil moisture time series from the NEON portal.

E1105: **Detection and estimation of local signals**
Presenter: **Xiao Fang**, The Chinese University of Hong Kong, Hong Kong

To segment a sequence of independent random variables at an unknown number of change-points, we introduce new procedures that are based on thresholding the likelihood ratio statistic or the maximum score statistic. We derive analytic approximations for the probability of a false positive error when there are no change-points, and justify some of the approximations rigorously.

E1137: **Novel network change point methods**
Presenter: **Ivor Cribben**, Alberta School of Business, Canada

Identifying change points in dynamic network structures has become increasingly popular across various domains, from neuroscience to telecommunication to finance. We will present two new change point detection methods. The first method uses non-negative matrix factorization, an unsupervised dimension reduction technique, and a new binary search algorithm to identify multiple change points. The second method considers changes in vine copula structure, various state-of-the-art segmentation methods to identify multiple change points, and a likelihood ratio test or the stationary bootstrap for inference. The vine copulas allow for various forms of dependence. We apply both methods to simulated, financial and to functional magnetic resonance imaging (fMRI) data sets.

EO376 Room Virtual R37 TIME SPACE MODELS: EVENTS AT RANDOM BEYOND GAUSSIANITY II Chair: Anastassia Baxevani

E1541: **Exceedance time distributions through the clipped Slepian process**
Presenter: **Henrik Bengtsson**, Lund University, Sweden

In physics and engineering literature, the distribution of the excursion time distribution for a stationary Gaussian process has been approximated by a stationary switch process with independently distributed switching times. The approach matched the covariance of the clipped Gaussian process with the one for the stationary switch process. We propose a similar but alternative matching of the expected value of the clipped Slepian process and the corresponding switch process initiated at zero. The method has several advantages over its original, stationary matching version. First, for a large class of processes, it produces a valid distribution for the excursion time, while the stationary version never leads to a valid distribution. Another advantage is when used for the non-zero excursion levels since it can utilize both the up-crossing Slepian model and the down-crossing one leading to natural approximations of the distributions of excursions above and below the given level.

E1544: **Multi-normex distributions**
Presenter: **Marie Kratz**, ESSEC Business School, CREAR, France
Co-authors: Evgeny Prokopenko

The purpose is to build a sharp approximation of the whole distribution of the sum of iid heavy-tailed random vectors, combining mean and extreme behaviors. It extends the so-called 'normex' approach from a univariate to a multivariate framework. We propose two possible multi-normex distributions, named d-Normex and MRV-Normex. Both rely on the Gaussian distribution for describing the mean behavior, via the CLT, while the difference between the two versions comes from using the exact distribution or the EV theorem for the maximum. The main theorems provide the rate of convergence for each version of the multi-normex distributions towards the distribution of the sum, assuming second-order regular variation property for the norm of the parent random vector when considering the MRV-normex case. Numerical illustrations and comparisons are proposed with various dependence structures on the parent random vector, using QQ-plots based on geometrical quantiles.

E1562: **Statistical learning for general point processes**
Presenter: **Ottmar Ottmar Cronie**, University of Gothenburg, Sweden

A first general (supervised) statistical learning framework is presented for point processes in general spaces, which is based on the combination of two new concepts: i) bivariate prediction errors, which are measures of discrepancy/prediction-accuracy between two-point processes, and ii) point process cross-validation (CV), which we define through point process thinning. The general idea is to carry out the fitting by predicting CV-generated validation sets using the corresponding training sets; the prediction error, which we minimise, is quantified through our bivariate prediction errors. Having presented some theoretical properties of our bivariate innovations, we look closer at the case where the CV procedure is obtained through independent thinning and we apply our statistical learning methodology to non-parametric kernel estimation of spatial intensity functions, showing numerically that it outperforms the state of the art in terms of mean (integrated) squared error. If time permits, we also highlight how our statistical learning approach can be applied to parametric intensity function estimation and Papangelou conditional intensity function estimation.

E1735: **Normal Pareto distributions: Theoretical framework and computational issues**
Presenter: **Tomasz Kozubowski**, University of Nevada Reno, United States
Co-authors: Matthew Ohemeng

A new, conditionally Gaussian, hierarchical stochastic model for heavy-tailed data is introduced which generalizes the Laplace probability distribution. We present some basic properties of this model and discuss related computational issues. We also briefly consider inferential aspects of the model.

EO122 Room Virtual R39 RECENT ADVANCEMENTS IN CAUSAL INFERENCE Chair: Joseph Antonelli

E0800: **Randomization tests for assessing covariate balance when designing and analyzing matched datasets**
Presenter: **Zach Branson**, Carnegie Mellon University, United States

Observational studies are often complicated by covariate imbalances among treatment groups, and matching methods alleviate this complication by finding subsets of treatment groups that exhibit covariate balance. Balance often serves as evidence that a matched dataset approximates a randomized experiment, but what kind of experiment does it approximate? We develop a randomization test to assess if matched data approximates a particular experimental design, such as complete randomization or block randomization. Our test can incorporate any design and allows for a graphical display that puts several designs on the same univariate scale, thereby allowing researchers to pinpoint which design, if any, is most appropriate for a matched dataset. After researchers determine a design, we recommend a randomization-based analytical approach that can incorporate any design and treatment effect estimator. We find that our test can frequently detect violations of randomized assignment, and also that matched datasets with high levels of balance tend to approximate balance-constrained designs like rerandomization, thereby allowing for precise causal analyses. However, assuming a precise design should be proceeded with caution, because it can harm inference if there are still large biases due to remaining imbalances after matching. We also demonstrate how this approach can be used for instrumental variable analyses and regression discontinuity designs, all using our R package randChecks.

E1000: **Statistical testing under distributional shifts: Applications in causal inference**
Presenter: **Niklas Pfister**, University of Copenhagen, Denmark

Statistical hypothesis testing is a central problem in empirical inference. Observing data from a distribution P, one is interested in testing whether P lies in a given null hypothesis while controlling the probability of false rejections. We will introduce a framework for statistical testing under distributional shifts. The goal will be to test a target hypothesis P in H_0 using observed data from a distribution Q, where we assume that P is related to Q through a known distributional shift. We propose a general testing procedure that first resamples from the observed data to construct an auxiliary data set (mimicking properties of P) and then applies an existing test in the target domain. We prove that this procedure holds pointwise asymptotic level if the target test holds pointwise asymptotic level, the size of the resample is at most of order $\sqrt{n}$, and the resampling weights are well-behaved. We will see that testing under distributional shifts naturally arises in causal inference and that the proposed procedure provides an easy-to-use and general-purpose solution to a wide variety of causal inference tasks.

E1068: **On the causal interpretation of randomized interventional indirect effects**
Presenter: **Caleb Miles**, Columbia University, United States

Natural indirect effects (NIEs) are mediated effects that can be identified when the exposure does not affect any confounders of the mediator-outcome relationship. To circumvent this assumption, so-called randomized interventional analog indirect effects (NIERs), which can be identified even in the presence of exposure-induced confounding, have gained popularity in the causal mediation literature. An essential property that a putative indirect effect must possess in order to have a true mediation/indirect effect interpretation is that it must be null whenever there is no one for whom both their exposure affects their mediator and their mediator affects their outcome. Without additional assumptions, the NIER does not satisfy this property. Further, examples will be provided of such additional assumptions under which this property can be recovered. Unfortunately, the NIE will also be identified under these additional assumptions, and so the NIER will provide little advantage over the NIE. Thus, while the NIER does have a meaningful interpretation pertaining to joint stochastic interventions on the exposure and intermediate variable, it cannot always be relied upon to tell stories about mediation.

E1722: **Integrated causal-predictive machine learning models for tropical cyclone epidemiology**
Presenter: **Rachel Nethery**, Harvard T.H. Chan School of Public Health, United States

Strategic preparedness reduces the adverse health impacts of hurricanes and tropical storms, referred to collectively as tropical cyclones (TCs), but its protective impact could be enhanced by a more comprehensive and rigorous characterization of TC epidemiology. To generate the insights and tools necessary for high-precision TC preparedness, we introduce a machine learning approach that standardizes estimation of historic TC health impacts, discovers common patterns and sources of heterogeneity in those health impacts, and enables identification of communities at the highest health risk for future TCs. The model integrates (1) a causal inference component to quantify the immediate health impacts of recent historic TCs at high spatial resolution and (2) a predictive component that captures how TC meteorological features and socioeconomic/demographic characteristics of impacted communities are associated with health impacts. We apply it to a rich data platform containing detailed historic TC exposure information and records of all-cause mortality and cardiovascular- and respiratory-related hospitalization among Medicare recipients. We report a high degree of heterogeneity in the acute health impacts of historic TCs, both within and across TCs, and, on average, substantial TC-attributable increases in respiratory hospitalizations. TC sustained windspeeds are found to be the primary driver of mortality and respiratory risks.

EO756 Room K2.31 Nash (Hybrid 07) MULTIVARIATE AND HIGH DIMENSIONAL TIME SERIES Chair: Sayar Karmakar

E0416: **A new classification method for multivariate time series data**
Presenter: **Soudeep Deb**, Indian Institue of Management Bangalore, India
Co-authors: Shubhajit Sen

Classification of multivariate time series (MTS) data has applications in various domains, for example, medical sciences, finance, sports analytics, etc. Though the classification of univariate time series (UTS) is a well-explored area, unfortunately, the same cannot be said for MTS classification. We propose a new technique that uses the advantages of dimension reduction through the t-distributed stochastic neighbor embedding (t-SNE) method, coupled with the attractive properties of the spectral density estimates of a time series, and k-nearest neighbor (k-NN) algorithm. We transform each MTS to a lower-dimensional time series, making it useful for visualizing and retaining the temporal patterns, and subsequently use that in classification. Then, we extend the standard univariate spectral density-based classification in the multivariate setting and prove its theoretical consistency. For real-life data analysis, we have chosen two health-related datasets. Empirically, at first, we establish that the pair-wise structure of the multivariate spectral density-based distance matrix is retained in the t-SNE transformed spectral density-based method. Then, for both cases, the proposed algorithm is implemented and we find that it achieves much better classification accuracy than the other widely used methods.

E0578: **Scalable Bayesian inference for time series via divide-and-conquer**
Presenter: **Deborshee Sen**, University of Bath, United Kingdom
Co-authors: David Dunson

Bayesian computational algorithms tend to scale poorly as data size increases. This had led to the development of divide-and-conquer-based approaches for scalable inference. These divide the data into subsets, perform inference for each subset in parallel, and then combine these inferences. While appealing theoretical properties and practical performance have been demonstrated for independent observations, scalable inference for dependent data remains challenging. We study the problem of Bayesian inference from very long time series. The literature focuses mainly on approximate approaches that lack any theoretical guarantees and may provide arbitrarily poor accuracy in practice. We propose a simple and scalable divide-and-conquer method, and provide accuracy guarantees. Numerical simulations and real data applications demonstrate the effectiveness of our approach.

E0586: **Bayesian vector autoregression using the tree rank prior**
Presenter: **Leo Duan**, University of Florida, United States

Vector autoregression is very popular for analyzing the multivariate time series. Besides good predictive performance, it enjoys nice interpretation in the Granger-causality graph — the past values of some variables are helpful for predicting the others. In the high dimensional setting with p variables, one often relies on the matrix-norm/matrix-rank based regularization to induce sparsity; however, this tends to create too many disconnected graph components that are difficult to interpret. To solve this problem, we propose a new type of low-rankness based on the graph topology — we define the "tree rank" as the number of spanning trees needed to cover the graph. Each spanning tree is a minimalist subgraph with $(p-1)$ edges but connects p nodes. As the result, having the regression coefficients on a few spanning trees leads to both high sparsity and high connectivity. To allow efficient computation and uncertainty quantification on the estimates, we develop a novel graph-based continuous shrinkage prior, that exploits a continuous relaxation for the spanning trees. This prior, which we call "Tree Rank Prior", avoids the costly combinatorial search in the graph estimation and enjoys the gradient-based Hamiltonian Monte Carlo algorithm for its posterior estimation. The model is applied to find the Granger causality graph in the functional magnetic resonance imaging data.

E1444: **A high dimensional Cramer-von Mises test**
Presenter: **Mengyu Xu**, University of Central Florida, United States

A Cramér-von Mises type test is developed for testing distributions of high dimensional continuous data and establish an asymptotic theory for quadratic functions of high-dimensional stochastic processes. To obtain cutoff values of our tests, we introduce two different procedures to

implement high-dimensional Cramér-von Mises test in practice: a plug-in calibration method and subsampling method. Theoretical justification and numerical studies of both approaches are provided.

EO655 Room K2.40 (Hybrid 08) MEDIATION ANALYSIS FOR COMPLEX DATA STRUCTURE (VIRTUAL) Chair: Yeying Zhu

E0814: **Causal mediation analysis based on partial linear models**
Presenter: **Xizhen Cai**, Williams College, United States
Co-authors: Yeying Zhu, Yuan Huang, Debashis Ghosh

The focus is on estimating the direct and indirect effects of mediation analysis based on a set of partial linear regression models. We allow a nonlinear relationship among the baseline covariates and the response variables in each model. Since we are only interested in estimating the coefficients for the treatment and the mediator in the structural models, we assume partial linear models where the baseline covariates are regarded as a nuisance. The estimates can be interpreted as causal effects without the linearity assumption. We also propose variable selection procedures when the set of mediators is high-dimensional. Simulation results show the superior performance of our proposed method and a data application is conducted when the set of candidate mediators are high-dimensional methylations.

E1030: **Longitudinal mediation analysis of time-to-event endpoints based on natural effect models**
Presenter: **Stijn Vansteelandt**, Ghent University and London School of Hygiene and Tropical Medicine, Belgium
Co-authors: Thang Tat Vo

The motivation comes from an analysis of the English Longitudinal Study of Ageing (ELSA), which aims to investigate the role of loneliness in explaining the negative impact of hearing loss on dementia. The methodological challenges that complicate this mediation analysis include the use of a time-to-event endpoint subject to competing risks, as well as the presence of feedback relationships between the mediator and confounders that are both repeatedly measured over time. To account for these challenges, we introduce natural effect proportional (cause-specific) hazard models. These extend marginal structural proportional (cause-specific) hazard models to enable effect decomposition. We show that under certain causal assumptions, the path-specific direct and indirect effects indexing this model are identifiable from the observed data. We next propose an inverse probability weighting approach to estimate these effects. On the ELSA data, this approach reveals little evidence that the total effect of hearing loss on dementia is mediated through the feeling of loneliness, with a non-statistically significant indirect effect equal to 1.012 (hazard ratio (HR) scale; 95% confidence interval (CI) 0.986 to 1.053).

E1512: **A doubly robust joint modelling approach of multiple uncausally correlated mediators**
Presenter: **Lijia Wang**, University of Waterloo, Canada
Co-authors: Yeying Zhu, Richard Cook

Causal mediation analysis aims at disentangling the effects of a treatment on an outcome via a variety of paths through either intermediate variables lied alongside the causal pathways (the mediators) or the treatment itself. Recently, mediation analysis on multiple mediators is attracting much attention because of inspirations from reality, where the relationship between the multiple mediators plays an important role when investigating the causal effects. Traditional studies focus on the scenario that the multiple mediators are causally sequentially related. We review and extend another new concept that the multiple mediators are uncausally related, which depicts the phenomenon that the multiple mediators are related given the baseline covariates but their correlation structure cannot be causally ordered or clearly identified. We further provide a copula-based joint modelling approach performing the causal mediation analysis of the multiple uncausally related mediators. A doubly robust approach is also proposed to tackle the model misspecification issue. Theoretical properties and simulation studies are also presented, with the theoretical standard error derived based on the sandwich formula. We finally apply the proposed method on a genetic psychiatric study dataset to identify the causal mediation effect of three DNA methylation loci on the pathway between childhood trauma and stress reactivity.

E1707: **Using the medflex package for (semi)high dimensional mediators and/or complex outcomes**
Presenter: **Theis Lange**, University of Copenhagen, Denmark

The goal of mediation analysis is typically not to assess mediation through a specific variable, but instead to assess mediation through a complex biomedical system. Such systems are typically measured through a range of variables each of which, only captures a reflection. The mediation analysis needs to take all such variables into account. It is demonstrated how this can be achieved using the medflex package for R.

EC851 Room K0.18 (Hybrid 03) CONTRIBUTIONS IN METHODOLOGICAL STATISTICS II Chair: Ingrid Van Keilegom

E1756: **Estimating treatment effects on optimal row designs under dependence**
Presenter: **Katerina Pericleous**, Cyprus University of Technology, Cyprus

The experimental units or simply units are arranged in time or along a line with every unit to be allocated one out of v treatments. The aim is to find the design which gives optimal estimates of treatments effects or of treatment differences. The main effects model with homogeneous population, when the observations follow a first-order autoregressive process, with positive or negative parameter p, is used. Universal optimality and other optimality are defined and shown that for positive p, the Williams IIa designs, which are A- and D-optimal for estimating treatment contrasts, are not A- or E-optimal for estimating treatment effects. In order to estimate treatment effects a shortened Williams design is applied by considering the first or last unit as the right alternative. In the case of three treatments and negative dependence, optimal designs are presented for any number of units.

E1657: **Optimal allocation strategies for blocked clinical trial designs with heterogeneous outcomes**
Presenter: **Lida Mavrogonatou**, University of Cambridge, United Kingdom
Co-authors: David Robertson, Sofia Villar

Clinical trials are multi-objective experiments, incorporating criteria such as power, type I error control, efficiency and patient benefit within and outside the trial. Depending on the trial at hand, optimal performance in terms of one or more of these criteria may take priority over others. The interrelation of the often-conflicting objectives is investigated in such cases. A blocked design setup is considered in which the variance of the treatment response error is allowed to vary both among treatments and blocks. It is shown that the balanced allocation no longer achieves optimal power or efficiency as in the homogeneous case and so alternative allocation strategies are examined. The D-optimal design is shown to be a promising allocation strategy, offering substantial gains in terms of power and estimation accuracy under variance heterogeneity.

E1362: **Mixed frontier estimation in the presence of measurement error with unknown variance**
Presenter: **Jun Cai**, KU Leuven, Belgium
Co-authors: Ingrid Van Keilegom

Stochastic frontier models for cross-sectional data typically assume that the one-sided distribution of firm-level inefficiency is either continuous or discrete. However, it may be reasonable to hypothesize that inefficiency is continuous except for a discrete mass at zero capturing fully efficient firms (zero-inefficiency). We extend a previous method for such a mixture distribution in the stochastic frontier model with unknown error variance and modify it to incorporate a lower bound frontier estimation as well. Consistency, convergence rates of the estimator are established, as well as a test of the zero-inefficiency hypothesis. Simulations and an application to the cost efficiency of US banks are provided.

E0506: **A unifying convex analysis framework to inference for penalized least squares**
Presenter: **Alberto Quaini**, University of Geneva, Switzerland

Co-authors: Fabio Trojani

A unifying convex analysis framework is proposed for studying the properties of a broad class of Penalized Least Squares Estimators (PLSEs) with convex penalties. The basis of such framework is a reinterpretation of PLSEs as proximity operators evaluated at a Least Squares Estimator. Asymptotically, these operators converge under standard assumptions to a limit proximity operator evaluated at a Gaussian random vector, which is uniquely characterized by the limit penalty of a PLSE. We apply these characterizations to study with a unified approach the asymptotic bias functional of PLSEs, Oracle properties of PLSEs, valid bootstrap approximations for PLSEs' asymptotic distribution, and PLSEs with singular designs.

EG065 Room Virtual R33 CONTRIBUTIONS IN COMPUTATIONAL AND METHODOLOGICAL STATISTICS Chair: Qing Wang

E1666: **Optimal repetitive reliability inspection of manufactured lots for lifetime models using prior information**

Presenter: **Carlos Perez-Gonzalez**, Universidad de La Laguna, Spain

Co-authors: Arturo J Fernandez, Vicent Giner-Bosch, Andres Carrion-Garcia

Repetitive group inspection of production lots is considered to develop the failure-censored plan with minimal expected sampling effort using prior information. Optimal reliability test plans are derived for the family of log-location and scale lifetime distributions, whereas a generalized beta distribution is assumed to model the nonconforming proportion, p. A highly efficient and quick step-by-step algorithm is determined in order to solve the underlying mixed nonlinear programming problem. Conventional repetitive group plans are often very effective in reducing the average sample number with respect to other inspection schemes, but sample sizes may increase under certain conditions such as high censoring. The inclusion of previous knowledge from past empirical results contributes to reducing drastically the amount of sampling required in life testing. Moreover, the use of expected sampling risks improves significantly the assessment of the real producer and consumer sampling risks. Several tables and figures are presented to analyze the effect of the available prior evidence about p. The results show that the proposed lot inspection scheme clearly outperforms the standard repetitive group plans obtained under the traditional approach based on conventional risks. Finally, an application to the manufacture of integrated circuits is included for illustrative purposes.

E0881: **A general framework for ID-based data structures to derive erm-approaches for learning**

Presenter: **Julian Gerstenberg**, Goethe University Frankfurt, Germany

An abstract ID-based data structure is defined as a contravariant functor from the category of finite sets with injections as morphisms to the category of measurable spaces. An abstract framework for ID-based data structures (ID = identifier) using the language of category theory is presented without assuming any knowledge of CT. Many important data structures are instances of the abstract definition: sequential data (data of this type is of the form $x_1, x_2, \ldots, x_n$ when using IDs $\{1, \ldots, n\}$), partitions, graphs, orders (total, partial,...), hierarchies, arrays and many more. Within the abstract framework, one can develop a rich exchangeability theory that is of foundational importance to many statistical applications, one of which being empirical risk minimization approaches for statistical learning tasks.

E1570: **Quantifying uncertainty of subsampling-based ensemble methods under a U-statistic framework**

Presenter: **Qing Wang**, Wellesley College, United States

Co-authors: Yujie Wei

The problem of variance estimation of subsampling-based ensemble methods, such as subbagging and sub-random forest, is addressed. We first recognize that a subsampling-based ensemble can be written in the form of a U-statistic of degree k, where k is the subsample size. As a result, one can study the uncertainty of the ensemble estimator under a U-statistic framework. Motivated by previous work, we propose to construct an unbiased variance estimator for a subsampling-based ensemble, which is efficient to realize with the help of a partition-resampling scheme. We show by simulation studies that the proposed variance estimator has a significant computational advantage and yields better performance in terms of mean, standard deviation, and mean squared error compared to the benchmark under either a simple linear regression model or a MARS model. Furthermore, we present how to construct an asymptotic confidence interval of the expected response of an ensemble using the proposed variance estimator, and compare its coverage probability with competing methods. In the end, we demonstrate the practical applications of the methodology using real data examples.

E1271: **Forecasting by learning the evolution-driving function**

Presenter: **Dalia Chakrabarty**, Brunel University London, United Kingdom

The state that a generic deterministic dynamical system is in, at any time in the future, is computable, as long as we have information on the function that drives the evolution of this system. Indeed, probabilistic learning of this evolution-driver will lead to probabilistic forecasting of states. In contrast, learning patterns in the already observed values of the phase space variables in the past, does not guarantee correct forecasting, irrespective of the sophistication of the learning and/or of parametrisation of information replication at the considered time point in the future. However, we do not possess any training data to permit supervised learning of the sought evolution driving function. So to enable such supervised learning, the evolution-driver is learnt at a given input, (namely, a given time and state), by embedding it in the support of the pdf of the phase space variables. This generates the originally-absent training set, using which we learn the sought function (by modelling it with a Gaussian Process), and predict it at the (future) test time. Phase space variables attained at that future time are then computed by inputting this forecast evolution-driving function using a generalised version of Newton's 2nd Law. An empirical illustration of the methodology is made to perform forecasting of daily new COVID19 infection numbers.

CI880 Room K E. Safra (Multi-use 01) FORECASTING (VIRTUAL) Chair: Michael Owyang

C0190: **On the real time predictive content of financial conditions for growth**

Presenter: **Michael McCracken**, Federal Reserve Bank of St. Louis, United States

Co-authors: Aaron Amburgey

Analytical, Monte Carlo, and empirical evidence is provided on the real-time predictive content of financial conditions indices, notably the NFCI, for quantiles of the distribution of U.S. real GDP growth. We do so by investigating two specific issues in the vulnerable growth literature. First, we construct (unofficial) real-time vintages of the NFCI. This allows us to conduct the out-of-sample analysis without introducing look-ahead biases that are naturally introduced when using a single current vintage. We then investigate the usefulness of asymptotic and bootstrap-based critical values for tests of predictive ability for nested models in the context of linear quantile regression. We find that for quantiles near the median, the asymptotic critical values can provide accurately sized tests in reasonable sample sizes. As the quantiles shift into the tails, the asymptotic critical values perform quite poorly even in large samples. Both the fixed regressor wild bootstrap and the tapered block bootstrap provide accurately sized tests across all quantiles even in modest samples.

C0194: **Forecasting sovereign defaults**

Presenter: **Ana Galvao**, University of Warwick, United Kingdom

Co-authors: Michael McCracken, Michael Owyang

Six countries defaulted partially in their debt obligations in 2020; the highest number of defaults recorded by Moodys-rated sovereign bonds since 1983. We propose a novel Panel Qualitative Vector Autoregressive (Qual-VAR) model to measure and forecast the probability of sovereign defaults. We use the sovereign default events recorded previously, a set of global factors (activity and financial factors) and country-specific macroeconomic variables (such as government debt, external debt and FDI inflows to GDP ratios and GDP growth) sampled quarterly from 1980 to 2020 to

estimate our proposed quantitative model for a panel of countries. Our panel of 50 countries includes emerging European, Asian, African, and Latin American economies. Using the panel Qual-VAR model, we can compute multi-step ahead forecasts for the probability of sovereign default for each country in our sample while allowing for serial correlation in the probability of default. The model is also useful to compute probability forecasts conditional on paths for the key global factors considered. Finally, the panel Qual-VAR allows us to assess the contribution of global versus country-specific factors in explaining sovereign default risks.

C0195: **Asymmetric loss and breakeven inflation forecasts**
Presenter: **Michael Owyang**, Federal Reserve Bank of St Louis, United States

Breakeven inflation rates (BEIs) are computed from the difference in yields on standard and inflation-protected Treasury securities. These BEIs are often interpreted as market-based inflation forecasts and used in empirical models as a measure of representative agent expectations. Previous studies, however, have shown that these market-based forecasts are not rationalizable for standard squared error loss functions. This result obtains, in part, because the BEIs are not unbiased. We reconsider the forecast rationality of BEIs using an alternative loss function that allows for asymmetric preferences.

CO278 Room Virtual R32 TRACKING THE ECONOMY WITH HIGH DIMENSIONAL METHODS Chair: Scott Brave

C1130: **Tracking U.S. consumers in real time with a new weekly index of retail trade**
Presenter: **Scott Brave**, Federal Reserve Bank of Chicago, United States

A new weekly index of retail trade is created that accurately predicts the U.S. Census Bureaus Monthly Retail Trade Survey (MRTS). The index weekly frequency provides an early snapshot of the MRTS and allows for more granular analysis of the aggregate consumer response to fast-moving events such as the Covid-19 pandemic. To construct the index, we extract the co-movement in weekly data series capturing credit and debit card transactions, foot traffic, gasoline sales, and consumer sentiment. To ensure that the index is representative of aggregate retail spending, we implement a novel benchmarking method that uses a mixed-frequency dynamic factor model to constrain the weekly index to match the monthly MRTS. We use the index to document several interesting features of U.S. retail sales during the Covid-19 pandemic, many of which are not visible in the MRTS. In addition, we show that our index would have more accurately predicted the MRTS in real-time during the pandemic when compared to either consensus forecasts available at the time, monthly autoregressive models, or other commonly-cited high-frequency data that aims to track retail spending. The gains are substantial, with roughly 50 to 75 percent reductions in mean absolute forecast errors.

C1131: **Subspace shrinkage in conjugate Bayesian vector autoregressions**
Presenter: **Florian Huber**, University of Salzburg, Austria
Co-authors: Gary Koop

Macroeconomists using large datasets often face the choice of working with either a large Vector Autoregression (VAR) or a factor model. We develop methods for combining the two using a subspace shrinkage prior. Subspace priors shrink towards a class of functions rather than directly forcing the parameters of a model towards some pre-specified location. We develop a conjugate VAR prior which shrinks towards the subspace which is defined by a factor model. The approach allows for estimating the strength of the shrinkage as well as the number of factors. After establishing the theoretical properties of our proposed prior, we carry out simulations and apply it to US macroeconomic data. Using simulations we show that our framework successfully detects the number of factors. In a forecasting exercise involving a large macroeconomic data set we find that combining VARs with factor models using our prior can lead to forecast improvements.

C1201: **A mixed frequency model for the euro area labour market**
Presenter: **Claudia Foroni**, European Central Bank, Germany
Co-authors: Agostino Consolo, Catalina Martinez Hernandez

We introduce a Bayesian Mixed-Frequency VAR model for the aggregate euro area labour market that features a structural identification via sign restrictions. The purpose is twofold: we aim at (i) providing reliable and timely forecasts of key labour market variables and (ii) enhancing the economic interpretation of the main movements in the labour market. We find satisfactory results in terms of forecasting, especially when looking at quarterly variables, such as employment growth and the job-finding rate. Furthermore, we look into the shocks that drove the labour market and macroeconomic dynamics from 2002 to early 2020, with a first insight also on the COVID-19 recession. While domestic and foreign demand shocks were the main drivers during the Global Financial Crisis, aggregate supply conditions and labour supply factors reflecting the degree of lockdown-related restrictions have been important drivers of key labour market variables during the pandemic.

C1522: **Sparse temporal disaggregation**
Presenter: **Luke Mosley**, Lancaster University, United Kingdom
Co-authors: Idris Eckley, Alex Gibberd

Temporal disaggregation is a method commonly used in official statistics to enable high-frequency estimates of key economic indicators, such as GDP. Traditionally, such methods have relied on only a couple of high-frequency indicator series to produce estimates. However, the prevalence of large, and increasing, volumes of administrative and alternative data-sources motivates the need for such methods to be adapted for high-dimensional settings. We propose a novel sparse temporal-disaggregation procedure and contrast this with the classical Chow-Lin method. We demonstrate the performance of the proposed method through a simulation study, highlighting various advantages realised. We also explore its application to disaggregation of UK gross domestic product data, demonstrating the method's ability to operate when the number of potential indicators is greater than the number of low-frequency observations.

CO224 Room Virtual R38 SPATIAL ECONOMETRICS AND STATISTICS FOR MICRO-GEOGRAPHIC DATA Chair: Diego Giuliani

C0260: **Director appointments, boardroom networks, and firm environmental performance**
Presenter: **Dakshina De Silva**, Lancaster University, United Kingdom
Co-authors: Aurelie Slechten, Mingyuan Chen

Using BoardEx (2000-2017), a dynamic network is created connecting firms and board directors for the United States. We use the Environmental Protection Agency's Toxic Release Inventory to measure environmental performance at the director and firm level. We examine how a candidate's past environmental performance and networks affect director appointments. This allows us to endogenize the effect of directors environmental experience when studying the impact on firms chemical releases. We show that firms are likely to appoint influential directors with good environmental records and similar characteristics. Further, boards with good environmental performance and with diverse environmental backgrounds improve firms' environmental performance.

C0261: **How highway expansion affect land use changes**
Presenter: **Jean Dube**, Universita Laval, Canada
Co-authors: Cedric Brunelle, Maroua Aikous

The expansion of transport infrastructure has important consequences for the spatial distribution and development of economic and residential activities. While much has been written on how such new infrastructure influences real estate prices, not much empirical investigation has been made about how it can influence the emergence of economic and commercial activities land use changes are scarcer. Using a highway extension project in the Montreal Metropolitan area quasi-natural experiment, it is investigated how the expansion of a highway influences the crowding-out

effect land use changes and businesses' location and the relocation related to the crowding-out effect within a relatively undeveloped suburb fringe for economic activities with intensive land consumption. To do so, a panel dataset of individual land parcels (or lots (or parcels) between 1995 and 2018 2019 is developed, while changes in land use are investigated according to key moments related to the construction and expansion of Highway 30 on Montreal's South Shore (Canada), which fully opened in 2012. Defining the treatment areas using the highway access ramp area, the analysis shows that the construction of new highway infrastructure may influence local economic activities land use changes over time by facilitating a crowding-out effect for activities that are important land consumers.

C0940: **The effect of agglomeration economies on firm deaths: A comparison of firm and regional based approaches**
Presenter: **Bernadette Power**, University College Cork, Ireland
Co-authors: Ryan Geraldine, Justin Doran

The merits of regional and firm based approaches for analyzing the effect of agglomeration economies on firm deaths in Ireland are compared. We aggregate a comprehensive dataset on Irish firm deaths to Electoral Division (ED) level, the lowest geographical scale available. Estimates of the effect of agglomeration on firm deaths from a regional analysis at ED level using cross-sectional spatial autoregressive techniques are compared to firm-level estimates from a contemporary log-log model with spatially weighted agglomeration regressors. While estimates of the effects of agglomeration using these alternative methods are much discussed in existing literature rarely are the approaches or results compared. We show that contrasting results are found using the same dataset dependent upon the unit of analysis used. Diversity lowers regional and firm deaths while specialization raises regional deaths but lowers firm deaths. Greater urbanization does not have a significant effect on firm hazards rates or equivalent regional estimates. While regional estimates provide evidence on the existence and nature of spatial dependence (positive in this case), firm estimates may provide evidence on whether this spatial dependence is due to diversity, specialization or urbanization economies. No empirical analysis to our knowledge directly compares regional and firm based approaches.

C1058: **Fitting approaches of Cliff-Ord models to data affected by locational errors**
Presenter: **Flavio Santi**, University of Verona, Italy
Co-authors: Diego Giuliani, Giuseppe Espa, Maria Michela Dickson

When a spatial regression model is fitted to micro-geographic data in order to account for spatial dependence, the quality of information on unit locations becomes relevant. Locational errors may originate from flaws in geocoding or georeferencing processes, as well as from geomasking of unit positions; as a result, the actual positions of units are known up to an error, whose probabilistic behaviour may be either known or unknown, and whose magnitude may be either homogeneous or heterogeneous amongst units. Spatial regression models a la Cliff-Ord are known to suffer heavily from the consequences of locational errors, as they typically make the parameter estimators markedly biased and inconsistent. A review is made for fitting approaches of Cliff-Ord models to data affected by locational errors. A new hybrid approach is proposed which combines analytical and computational methods.

CO665 Room Virtual R40 ASSET PRICING I Chair: Julien Penasse

C1492: **Dynamic asset (mis)pricing: Build-up vs. resolution anomalies**
Presenter: **Jules Van Binsbergen**, Wharton and NBER, United States

Asset pricing anomalies are classified into those that exacerbate mispricing (build-up anomalies) and those that resolve them (resolution anomalies). To this end, we estimate the dynamics of price wedges for a large number of well-known anomaly portfolios in the factor zoo and map them to firm-level mispricings. We find that several prominent anomalies like momentum and profitability further dislocate prices. While mispricing buildup is often quick, the subsequent resolution tends to be slow, suggesting the potential for material real economic consequences. The results suggest that financial intermediaries chasing build-up anomalies in fact negatively affect price efficiency and associated real capital allocation.

C1509: **Model complexity, expectations, and asset prices**
Presenter: **Andrea Vedolin**, Boston University, United States

The purpose is to analyze how limits to the complexity of statistical models used by market participants can shape asset prices. We consider an economy in which agents can only entertain models with at most k factors, where k may be distinct from the true number of factors that drive the economy's fundamentals. We first characterize the implications of the resulting departure from rational expectations for return dynamics and relate the extent of return predictability at various horizons to the number of factors in the agent's models and the statistical properties of the underlying data-generating process. We then apply our framework to two applications in asset pricing: (i) violations of uncovered interest rate parity at different horizons and (ii) momentum and reversal in equity returns. We find that constraints on the complexity of agents models can generate return predictability patterns that are consistent with the data.

C1559: **Smart stochastic discout factors**
Presenter: **Fabio Trojani**, University of Geneva, University of Turin and SFI, Switzerland
Co-authors: Alberto Quaini

A novel no-arbitrage framework is proposed which exploits convex asset pricing constraints to study the properties of investors marginal utility of wealth or, more generally, Stochastic Discount Factors (SDFs). We establish a duality between minimum dispersion SDFs and suitable penalized portfolio selection problems, building the foundation for a nonparametric characterization of the feasible tradeoffs between an SDFs pricing accuracy and its comovement with systematic risks. Empirically, we find that a minimum variance correction of a CAPM SDF produces a Pareto optimal tradeoff. This Pareto optimal SDF only depends on two economically distinct risk factors: A market factor and a minimum variance excess return factor, which optimally bounds the aggregate mispricing of risks unspanned by market risk.

C0171: **Discussant**
Presenter: **Julien Penasse**, University of Luxembourg, Luxembourg

Based on the session speakers' recent work, relevant progress in the field of sampling for large-scale data will be discussed.

CC859 Room K2.41 (Hybrid 09) CONTRIBUTIONS IN FINANCIAL ECONOMETRICS II Chair: Genevieve Gauthier

C1283: **The leverage effect and propagation**
Presenter: **Leopoldo Catania**, Aarhus BBS, Denmark

A new way to measure the leverage effect and its propagation over time is proposed. We show that, with respect to the newly proposed measure, common volatility models like the GJRGARCH, the Exponential GARCH, and the asymmetric SV can be inaccurate to correctly represent the leverage effect and its propagation for financial time series. We propose to modify the variance recursion of common volatility models by including an auxiliary leverage process which allows for a proper representation of the leverage effect and its propagation over time. Empirical results indicate that the inclusion of the auxiliary leverage process is required for both in-sample and out-of-sample analyses.

C1289: **A penalized two-pass regression to predict stock returns with time-varying risk premia**
Presenter: **Gaetan Bakalli**, Auburn university, United States
Co-authors: Stephane Guerrier, Olivier Scaillet

A penalized two-pass regression with time-varying factor loadings is developed. The penalization in the first pass enforces sparsity for the time-variation drivers while also maintaining compatibility with the no-arbitrage restrictions by regularizing appropriate groups of coefficients. The

second pass delivers risk premia estimates to predict equity excess returns. Our Monte Carlo results and our empirical results on a large cross-sectional data set of US individual stocks show that penalization without grouping can yield to nearly all estimated time-varying models violating the no-arbitrage restrictions. Moreover, our results demonstrate that the proposed method reduces the prediction errors compared to a penalized approach without appropriate grouping or a time-invariant factor model.

C0426: **Fiscal policy, international spillovers, and endogenous productivity**
Presenter: **Mathias Klein**, Sveriges Riksbank, Sweden

Empirical evidence is presented on the international effects of US fiscal policy from structural vector autoregressions identified through external instruments in a panel setting for the G7 countries. An exogenous increase in US government spending is estimated to produce sizeable positive responses of output and consumption in the rest of the G7 countries, both about half as large as their domestic US counterparts, while strongly depreciating the US terms of trade and lowering short-run real interest rates. Moreover, fiscal shocks are estimated to have a strongly positive impact on hourly labor productivity in the private sector. We solve a two-country New Keynesian model in closed form and show that a low-cost elasticity of varying technology utilization can simultaneously explain the positive productivity, consumption and international spillover effects as well as the real depreciation resulting from expansionary US government spending shocks.

C0269: **Venturing into unchartered territories: An extensible parametric implied volatility surface model**
Presenter: **Genevieve Gauthier**, HEC Montreal, Canada
Co-authors: Pascal Francois, Remi Galarneau-Vincent, Frederic Godin

A new parametric representation of implied volatility surfaces is proposed. The factors adequately capture the moneyness and maturity slopes, the smile attenuation, and the smirk. Furthermore, the implied volatility specification is twice continuously differentiable and well behaved asymptotically, allowing for clean interpolation and extrapolation over a wide range of moneyness and maturity. Fitting performance on S&P 500 options compares favourably with existing benchmarks. The benefits of a smoothed implied volatility surface are illustrated through the valuation of illiquid index derivatives, the extraction of the risk-neutral density and risk-neutral moments, the calculation of option price sensitivities, and the calculation of SVIX for the equity risk premium lower bound.

CG015 Room Virtual R35 CONTRIBUTIONS IN MACHINE LEARNING FOR ECONOMETRICS AND FINANCE Chair: Marie Bessec

C1538: **Binary choice with asymmetric loss in a data-rich environment: theory and an application to racial justice**
Presenter: **Andrii Babii**, UNC Chapel Hill, United States
Co-authors: Eric Ghysels

The binary choice problem in a data-rich environment with asymmetric loss functions is studied. In contrast to asymmetric regression problems, the binary choice with general loss functions and high-dimensional datasets is challenging and not well understood. Econometricians have studied non-parametric binary choice problems for a long time, but the literature does not offer computationally attractive solutions in data-rich environments. In contrast, the machine learning literature has many algorithms that form the basis for much of the automated procedures that are implemented in practice, but is focused mostly on loss functions that are independent of individual characteristics. We show that the theoretically valid predictions of binary outcomes with a generic loss function can be achieved via a very simple reweighting of the logistic regression or state-of-the-art machine learning techniques, such as LASSO, boosting, or deep learning. We apply our analysis to racial justice in pretrial detention.

C1360: **A green wave in media, a change of tack in stock markets**
Presenter: **Marie Bessec**, University Paris Dauphine, France
Co-authors: Julien Fouquau

The impact of green sentiment in US media on financial markets is explored. Using textual analysis with a dictionary-based approach, we retrieve several scores of attention, tonality and uncertainty in the coverage of environmental news of four major US newspapers. We consider various weighting schemes to account for the visibility and relevance of the text sources and several sets of newspapers to measure the possible impact of their editorial line. The results establish that greater attention to environmental news in US media reduced the excess returns of carbon-intensive stocks and increased their volatility over the last decade, especially when the coverage was negative or uncertain. The opposite result holds for the most virtuous green assets. Restricting the corpus of texts to conservative newspapers mitigates the impact of the coverage. Overall, our findings illustrate how rising environmental concerns lead investors to shift their asset allocation.

C1383: **Predicting emerging market credit spreads with support vector regression: Exploiting ubiquitous local optima**
Presenter: **Gary Anderson**, CEMAR LLC, United States
Co-authors: Alena Audzeyeva

A coherent framework is proposed using support vector regression (SVR), for generating and ranking a set of high-quality models for predicting emerging market sovereign credit spreads. Our framework adapts a global optimization algorithm employing an hv-block cross-validation metric, pertinent for models with serially correlated economic variables, to produce robust sets of tuning parameters for SVR kernel functions. In contrast to previous approaches identifying a single best tuning parameter setting, a task that is practically unattainable in many financial market applications, we proceed with a collection of tuning parameter candidates, employing the model confidence set test to select the most accurate models from the collection of promising candidates. Using bond credit spread data for three large emerging market economies and an array of input variables motivated by economic theory, we apply our framework to identify relatively small sets of SVR models with superior out-of-sample forecasting performance. Benchmarking our SVR forecasts against the random walk and conventional linear model forecasts provides evidence for the notably superior forecasting accuracy of SVR-based models. Consequently, our evidence indicates a better ability of highly flexible SVR to capture investor expectations about future spreads reflected in today's credit spread curve.

C0221: **Real estate pricing and natural language processing: Improving price valuation using textual descriptions**
Presenter: **Katharina Baur**, Albert-Ludwigs-Universitat, Germany

A text mining approach is suggested for the valuation of residential properties. Investment decisions in the real estate sectors are generally of long-term nature. Accurate and fair pricing is, therefore, fundamental to prevent miscalculations of such investments. As the focus of previous research lies on price predictions with solely numerical features, a new approach is addressed by including written descriptions about real estate properties for assessing its value. Essentially, we make use of state-of-the-art natural language processing methods, namely contextualized embeddings, to employ textual information. We show that prediction models can significantly benefit from these methods compared to standard approaches.

Monday 20.12.2021 16:55 - 18:35 Parallel Session R – CFE-CMStatistics

EO124 Room K0.16 (Hybrid 02) ADVANCES IN THE ANALYSIS OF DEPENDENT FUNCTIONAL DATA STRUCTURES Chair: Anne van Delft

E0466: **Semiparametric functional factor models with Bayesian Rank Selection**
Presenter: **Daniel Kowal**, Rice University, United States
Co-authors: Antonio Canale

Functional data are frequently accompanied by parametric templates that describe the typical shapes of the functions. Although the templates incorporate critical domain knowledge, parametric functional data models can incur significant bias, which undermines the usefulness and interpretability of these models. To correct for model misspecification, we augment the parametric templates with an infinite-dimensional nonparametric functional basis. Crucially, the nonparametric factors are regularized with an ordered spike-and-slab prior, which implicitly provides consistent rank selection and satisfies several appealing theoretical properties. This prior is accompanied by a parameter-expansion scheme customized to boost MCMC efficiency, and is broadly applicable for Bayesian factor models. The nonparametric basis functions are learned from the data, yet constrained to be orthogonal to the parametric template in order to preserve distinctness between the parametric and nonparametric terms. The versatility of the proposed approach is illustrated through applications to synthetic data, human motor control data, and dynamic yield curve data. Relative to parametric alternatives, the proposed semiparametric functional factor model eliminates bias, reduces excessive posterior and predictive uncertainty, and provides reliable inference on the effective number of nonparametric terms–all with minimal additional computational costs.

E0654: **Estimating the conditional distribution in functional regression problems**
Presenter: **Siegfried Hoermann**, Graz University of Technology, Austria
Co-authors: Gregory Rice, Thomas Kuenzer

The problem of consistently estimating the conditional distribution of a functional data object given covariates in a general space is considered. Thereby we assume that the response and the covariate are related by a linear regression model. Two natural estimation methods are proposed, based on either bootstrapping the estimated model residuals, or fitting functional parametric models to the model residuals and estimating the conditional distribution via simulation. Whether either of these methods lead to consistent estimators depends on the consistency properties of the regression operator estimator, and the space within which the response is viewed. We show that under general consistency conditions on the regression operator estimator, consistent estimation of the conditional distribution can be achieved, both when the response is an element of a separable Hilbert space, and when it is an element of the Banach space of continuous functions. The latter result implies that we can estimate the conditional probability of certain path properties, which are of interest in applications. The proposed methods are studied in several simulation experiments, and data analyses of electricity price and pollution curves. We also demonstrate how our method can be used for constructing confidence regions and in the context of functional quantile regression.

E0662: **Sparsely observed functional data on the sphere**
Presenter: **Alessia Caponera**, EPFL, Switzerland
Co-authors: Julien Fageot, Matthieu Simeoni, Victor Panaretos

Asymptotic theory for sparsely observed functional data has been largely developed when the domain is the interval $[0,1]$, in both the i.i.d. and time-dependent settings. For instance, the covariance/autocovariance functions can be suitably estimated by local polynomials, which are well defined in a 2-dimensional planar domain. However, when the domain is the sphere, the task consists in estimating a function on a 4-dimensional non-flat surface (i.e., $\mathbb{S}^2 \times \mathbb{S}^2$) and it could be convenient to consider other methods which naturally incorporates such structure. To this purpose, we define our estimator as the minimizer of a Tikhonov regularization problem. We hence make use of the machinery of reproducing kernel Hilbert spaces and specifically spherical Sobolev spaces to give a full characterization of the solution. The main result consists of an optimal rate of convergence for the covariance function estimator which can be interpreted in both the dense and sparse regimes. Additionally, we provide the extension to the stationary time-series framework, thus considering the autocovariance functions at different lags. The findings are validated through numerical experiments.

E0910: **Persistence surfaces: A new frequency specific topological summary for time series dependence structure**
Presenter: **Anass El Yaagoubi Bourakna**, King Abdullah University of Science and Technology, Saudi Arabia
Co-authors: Hernando Ombao, Moo K Chung

Topological data analysis (TDA) has become a powerful approach over the last years, mainly because of its ability to capture the shape present in the geometry of the data at hand. However, TDA methods that rely on Rips or Morse filtrations produce highly variable results due to sensitivity to noise, regardless of the stability theorems that have been derived, which are too restrictive to be of practical importance when it comes to time series analysis. In many applications such as brain signals, the raw form of the data does not always display any geometrical form or shape. Nevertheless, by transforming the raw data into connectivity matrices, some geometrical information might be present in the underlying connectivity network, even if no apparent geometrical structure is present in the raw form of the data. We propose a new frequency-specific topological summary for analyzing time-series data. This new topological summary that we call persistence surface (PS) can be viewed as an extension of the popular persistence landscape (PL). All the statistical results that have been derived for the PL will hold for the PS due to its definition. We demonstrate that our approach is able to detect topological features in simulated data and provides new insights on existing datasets of brain signals using functional data analysis.

EO446 Room K0.18 (Hybrid 03) RECENT ADVANCES IN BAYESIAN MODELLING Chair: Bruno Santos

E0794: **Spatial multi-resolution model for forestry data**
Presenter: **Isa Marques**, University of Goettingen, Germany
Co-authors: Paul Wiemann, Thomas Kneib

Geophysical processes often yield datasets that are spatially irregular, broadcasting a multi-scale character over space where observations are collected at different intensities in different areas. In such cases, models that assume the same dependence structure over the whole space - single-resolution stationary spatial models - can be inappropriate. This is the case for many forestry datasets, in which data is intensively collected in several distanced and relatively small-sized plots. In such cases, within plot data is typically characterized by large spatial ranges, while the spatial range between plots is relatively smaller. We develop a Bayesian spatial multi-resolution technique that models separately local and global processes, while allowing dimensions to interact. The resulting model is non-stationary and performs well for small local datasets, as well as for generally large datasets, through the use of Gaussian Markov random fields. The performance of our model is compared to that of standard single and multi-resolution models, for both simulated and real datasets. The resulting Bayesian model can be extended to fit spatio-temporal data.

E0924: **Semi-implicit variational inference in additive models**
Presenter: **Jens Lichter**, University of Goettingen, Germany
Co-authors: Paul Wiemann, Thomas Kneib

For big-scale problems and complex Bayesian regression models, variational inference (VI) offers a computationally efficient way of approximating the posterior distribution when no analytic form exists. Classical VI, however, is often based on the strong mean-field assumption, where, in the approximation, parameters are assumed to be independent of each other. As a consequence, parameter uncertainties are often underestimated. This

issue appears, in particular, in regression models with strongly correlated covariates. We propose to use the semi-implicit VI (SIVI) approach in additive models as an inferential method to weaken the mean-field assumption and improve uncertainty estimation. Firstly, SIVI uses a hierarchical construction of the parameters to restore parameter dependencies. Secondly, the mixing distribution on the higher level of the hierarchy does not need to be explicit, meaning a highly flexible implicit distribution represented by a neural network can be chosen. We present results from a simulation study revealing that SIVI accurately estimates parameter uncertainties and can outperform classical VI in additive models. Furthermore, we demonstrate our approach with an application to tree height models on a large-scale forestry data set.

E1363: **A defective cure rate quantile regression model for a maternal population with severe COVID-19**
Presenter: **Agatha Rodrigues**, Universidade Federal do Espirito Santo, Brazil
Co-authors: Patrick Borges, Bruno Santos

The aim is to address the problem of assessing the age and ethnicity on the specific survival times of pregnant and postpartum women hospitalized with severe acute respiratory syndrome confirmed by COVID-19 when cure is a possibility, where there is also the interest of explaining this impact on different quantiles of the survival times. To this end, we fitted a quantile regression model for survival data in the presence of long-term survivors based on the generalized distribution of Gompertz in a defective version, which is conveniently reparametrized in terms of the q-th quantile and then linked to covariates via a logarithm link function. The considered approach allows us to obtain how each variable affects the survival times in different quantiles. In addition, we are able to study the effects of covariates on the cure rate as well. We consider Markov Chain Monte Carlo (MCMC) methods to develop a Bayesian analysis in the proposed model and we evaluate its performance through a Monte Carlo simulation study. The study is part of the Brazilian Obstetric Observatory, a multidisciplinary project that aims to monitor and analyze public data from Brazil in order to disseminate relevant information in the area of maternal and child health.

E1285: **Growth curves for multiple-output response variables via Bayesian quantile regression models**
Presenter: **Bruno Santos**, University of Kent, United Kingdom
Co-authors: Agatha Rodrigues, Thomas Kneib

Reference fetal growth curves play an important role in identifying fetal growth restriction, macrosomia and other fetal malformations. This is verified based on percentiles of some biometric measurements at a specific gestational age using obstetric ultrasound. As an example, the diagnosis of microcephaly is based on a biparietal diameter smaller than the 10th percentile based on the reference curve. In practice, each biometric measurement reference curve is constructed independently of other measurements, even if they are correlated and some information about dependencies among them might be lost. Here we use these measurements to define growth curves modelling jointly more than one measurement. We consider structured additive quantile regression models for multiple-output response variables, where we are able to specify a nonlinear effect of time. We define a Markov Chain Monte Carlo (MCMC) procedure for model estimation, using ideas previously discussed in the literature. We examine four different ultrasound measurements and we show how one can retrieve more information when modelling these response variables jointly instead of individually. We illustrate the method with data from pregnancies from the University Hospital of the University of Sao Paulo (HU / USP) in the city of Sao Paulo, Brazil.

EO228 Room K0.19 (Hybrid 04) STATISTICS IN NEUROSCIENCE II Chair: Russell Shinohara

E0230: **Longitudinal ComBat: A method for harmonizing longitudinal multi-scanner imaging data**
Presenter: **Kristin Linn**, University of Pennsylvania, United States
Co-authors: Russell Shinohara, Joanne Beer

Neuroimaging is a major underpinning of modern neuroscience research and the study of brain development, abnormality, and disease. Combining neuroimaging datasets from multiple sites and scanners can increase statistical power for detecting biological effects of interest. However, technical variation due to differences in scanner manufacturer, model, and acquisition protocols may bias estimation of these effects. Originally proposed to address batch effects in genomic datasets, ComBat has been shown to be effective at removing unwanted variation due to scanner in cross-sectional neuroimaging data. We propose an extension of the ComBat model for longitudinal data and demonstrate its performance using simulations as well as longitudinal cortical thickness data from the Alzheimers Disease Neuroimaging Initiative (ADNI) study. We demonstrate that longitudinal ComBat controls type I error and has higher power for detecting changes in thickness over time compared to naively applying cross-sectional ComBat to the longitudinal trajectories.

E1049: **Statistical harmonization for the neuroimaging data with complex data distributions**
Presenter: **Haochang Shou**, University of Pennsylvania, United States
Co-authors: Andrew Chen, Russell Shinohara

With the increasing need for big data analytics in medical imaging, pooling and integrating data from multi-site studies has become critical. Site differences attributed to various sources are known to exist and might result in a substantial impact on the analytic results. Recently, batch-effect correction methods such as ComBat and CovBat have been successfully adapted to remove scanner and site differences in multimodal neuroimaging data and applied in many large-scale studies. However, the model assumptions of the existing statistical harmonization methods might restrict their direct application to broader imaging modalities with complex distributions. For example, fewer methods are available to harmonize the resting-state functional magnetic resonance imaging (fMRI) connectivity matrices, given the complex dependency structures temporally and spatially in the raw fMRI data and that the derived connectivity matrices do not necessarily belong to Euclidean metric space. Additionally, the current ComBat or CovBat assumes a Gaussian residual error and does not apply to imaging measures with skewed distribution or zero abundance such as white matter lesion counts. We will discuss several extensions of the statistical harmonization methods to complex neuroimaging modalities including multisite functional connectivity data and white matter hyperintensity data. The methods are shown to promote a more robust community detection in network analysis.

E1446: **Approximate hidden Semi-Markov models for dynamic connectivity analysis in resting-state fMRI**
Presenter: **Mark Fiecas**, University of Minnesota, United States

Motivated by a study on adolescent mental health, we conduct a dynamic connectivity analysis using resting-state functional magnetic resonance imaging (fMRI) data. A dynamic connectivity analysis investigates how the interactions between different regions of the brain, represented by the different dimensions of a multivariate time series, change over time. Hidden Markov models (HMMs) and hidden semi-Markov models (HSMMs) are common analytic approaches for conducting dynamic connectivity analyses. However, existing approaches for HSMMs are limited in their ability to incorporate covariate information. We approximate an HSMM using an HMM for modeling multivariate time series data. The approximate HSMM model allows one to explicitly model dwell-time distributions that are available to HSMMs, while maintaining the theoretical and methodological advances that are available to HMMs. We conducted a simulation study to show the performance of the approximate HSMM relative to other approaches. Finally, we used the approximate HSMM to conduct a dynamic connectivity analysis, where we showed how dwell-time distributions vary across the severity of non-suicidal self-injury (NSSI) in female adolescents.

E1708: **Tumor radiogenomics in gliomas with Bayesian layered variable selection**
Presenter: **Veera Baladandayuthapani**, University of Michigan, United States

A statistical framework is proposed to integrate radiological magnetic resonance imaging (MRI) and genomic data to identify the underlying radiogenomic associations in lower-grade gliomas (LGG). We devise a novel imaging phenotype by dividing the tumor region into concentric spherical layers that mimics the tumor evolution process. MRI data within each layer is represented by voxel–intensity-based probability density

functions which capture the complete information about tumor heterogeneity. Under a Riemannian-geometric framework, these densities are mapped to a vector of principal component scores which act as imaging phenotypes. Subsequently, we build Bayesian variable selection models for each layer with the imaging phenotypes as the response and the genomic markers as predictors. Our novel hierarchical prior formulation incorporates the interior-to-exterior structure of the layers, and the correlation between the genomic markers. We employ a computationally-efficient Expectation–Maximization-based strategy for estimation. Simulation studies demonstrate the superior performance of our approach compared to other approaches. With a focus on the cancer driver genes in LGG, we discuss some biologically relevant findings. Genes implicated with survival and oncogenesis are identified as being associated with the spherical layers, which could potentially serve as early-stage diagnostic markers for disease monitoring, prior to routine invasive approaches.

EO238 Room K0.50 (Hybrid 06) ANALYSIS OF DATA FROM WEARABLE DEVICES (VIRTUAL) Chair: Jaroslaw Harezlak

E0614: **Estimation of sparse functional quantile regression with measurement error: A SIMEX approach**
Presenter: **Carmen Tekwe**, Indiana University - Bloomington, United States

Quantile regression is a semiparametric method used to model associations between variables. It is most helpful when the covariates have a complex relationship with the location, scale, and shape of the outcome distribution. Despite its robustness to distributional assumptions and outliers in the outcome, regression quantiles may be biased in the presence of measurement error in the covariates. The impact of function-valued covariates contaminated with heteroscedastic error has not yet been examined; although, studies have investigated the case of scalar-valued covariates. Here, we present a two-stage strategy to consistently fit linear quantile regression models with a function-valued covariate that may be measured with error. In the first stage, an instrumental variable is used to estimate the covariance matrix associated with the measurement error. In the second stage, simulation extrapolation (SIMEX) is used to correct the measurement error in the function-valued covariate. Point-wise standard errors are estimated by means of nonparametric bootstrap. We present simulation studies to assess the robustness of the measurement-error-corrected for functional quantile regression. Our methods are applied to National Health and Examination Survey data to assess the relationship between physical activity and body mass index among adults in the United States.

E0412: **Modeling accelerometer-based physical activity**
Presenter: **Loki Natarajan**, University of California San Diego, United States

The use of wearable sensors to monitor physical activity is ubiquitous. These devices provide activity measurements at the minute-level (or even 30 Hz-level). Extracting useful information and modelling these densely sampled data can be challenging. We discuss functional data and other methods for analyzing accelerometer-derived physical activity measurements. We implement functional principal component analysis to study temporal patterns and derive the main modes of variation in accelerometer-based physical activity data and functional regression to estimate associations between functional physical activity profiles and health outcomes in a longitudinal setting. We discuss methodological issues pertinent to nested functional data (e.g., multiple visits or days of accelerometer wear within a participant). We illustrate methods via simulations and through application to a longitudinal study of physical activity and glucoregulatory markers.

E0842: **Functional survey weighted modeling to associate physical activity and non-alcoholic fatty liver disease**
Presenter: **Ekaterina Smirnova**, Virginia Commonwealth University, United States

Accelerometers present an objective alternative to assessing physical activity in multiple settings and allow continuous monitoring of physical activity in both lab and free-living environments. The National Health Examination Study (NHANES) is the largest US-population study that contains publicly available physical activity data obtained from wearable accelerometers together with extensive health and occupational information. The raw data is typically summarized into a minute-level accelerometry count measure, which leads to functional data collected over 1440 minutes per subject day. The NHANES study participants are recruited from the US population according to a survey design procedure, which has to be accounted for in statistical inference. To accurately associate these data with health outcomes, day-level summaries such as total activity volume, time spent in total sedentary, light and moderate to vigorous activities are often used. However, these summaries may be highly correlated with each other and may not capture the full complexity of the functional data patterns. We illustrate the utility of traditionally summarized features and functional data models that account for the complex survey design in the context of predicting the development and progression of non-alcoholic fatty liver disease (NAFLD). We further discuss the connection of physical activity to racial, occupational and socio-economic disparities in patients with NAFLD.

E1023: **A two-stage model for wearable device data**
Presenter: **Jiawei Bai**, Johns Hopkins University, United States
Co-authors: Jiawei Bai, Yifei Sun, Jennifer Schrack, Ciprian Crainiceanu, Mei-Cheng Wang

Recent advances in wearable computing technology have allowed continuous health monitoring in large observational studies and clinical trials. Examples of data collected by wearable devices include minute-by-minute physical activity proxies measured by accelerometers or heart rate. The analysis of data generated by wearable devices has so far been quite limited to crude summaries, for example, the mean activity count over the day. To better utilize the full data and account for the dynamics of activity level in the time domain, we introduce a two-stage regression model for the minute-by-minute physical activity proxy data. The model allows for both time-varying parameters and time-invariant parameters, which helps capture both the transition dynamics between active/inactive periods (Stage 1) and the activity intensity dynamics during active periods (Stage 2). The approach extends methods developed for zero-inflated Poisson data to account for the high dimensionality and time-dependence of the high-density data generated by wearable devices.

EO100 Room Virtual R18 STATISTICAL METHODS FOR PROVIDER PROFILING Chair: Michael Daniels

E0475: **Defining and estimating reliability in hierarchical logistic regression models for health care provider profiling**
Presenter: **Susan Paddock**, NORC at the University of Chicago, United States
Co-authors: John Adams, Jessica Hwang

In health care provider profiling, the reliability of a performance measure indicates whether observed differences in patient outcomes can be attributed to genuine differences in quality across providers. While reliability is easy to define, estimate, and interpret when the outcome of interest is continuous, and a hierarchical linear model can be assumed, several different definitions and estimators of reliability are in use for performance measures based on binary outcomes. We compare these candidate definitions and estimators when a hierarchical logistic regression model is assumed for the binary outcome. The salient differences between various definitions are demonstrated in simulations and on a data set of Florida primary care physicians treating Medicare fee-for-service beneficiaries.

E0610: **High-dimensional fixed effects profiling models: New developments and applications**
Presenter: **Danh Nguyen**, University of California, Irvine, United States

Profiling analysis aims to evaluate health care providers, such as hospitals, nursing homes, or dialysis facilities etc., with respect to a patient 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. We illustrate the versatility of FE profiling models 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 adverse recurrent events, and (c) new insights on operating characteristics such performance of FE model under the low information context/sparse outcome data setting.

E1154: **Individualized empirical null for profiling healthcare providers**
Presenter: **Kevin He**, University of Michigan, United States

Existing methods for healthcare provider profiling typically assume that the risk adjustment is perfect and the between-provider variation is entirely due to the quality of care. However, in practice, even with very good models for risk adjustment, there will be characteristics of patients and perhaps providers that are not completely accounted for (e.g. unobserved socio-economic factors and comorbidities), and many of these characteristics will be related to the outcome and vary across providers. Thus, some of the between-provider variation in a quality measure will typically be due to this incomplete risk adjustment (or unmeasured confounders), which should be recognized in assessing and monitoring providers. Otherwise, conventional methods disproportionately identify larger providers, although they need not be "extreme". To fairly assess providers, we propose an individualized empirical null method that accounts for the unexplained variation between providers. The proposed method robustly models the between-provider variance as a function of effective provider size and avoids bias against large providers.

E1480: **Quantitative bias analysis for provider profiling and practice variation studies**
Presenter: **Rolf Groenwold**, Leiden University Medical Center, Netherlands

Provider profiling and comparisons of performance indicators across centres essentially focus on understanding the (causal) effect of visiting one institution instead of another. At the same time, practice variation (e.g., in the use of medical treatment across centres) could provide a starting point for studies of the effects of medical treatments. Both instances require clear articulation of the assumptions needed to give the results of the analyses a causal interpretation. Violations of those assumptions (for example, unmeasured confounding or lack of conditional exchangeability) thus impacts provider profiling, just as it would impact observational studies of medical treatments. Unfortunately, commonly used methods are not immune to violations of the assumptions and the impact of that could be investigated using quantitative bias analysis methods. The possible role of quantitative bias analysis will be discussed for provider profiling studies and for observational studies that utilize practice variation.

EO072 Room Virtual R20 STATISTICAL METHODS FOR CONTEMPORARY BUSINESS APPLICATION Chair: Gourab Mukherjee

E1200: **A Bayesian structural uncertainty model to target rebates to consumers with correlated preferences**
Presenter: **Sivaramakrishnan Siddarth**, Marshall School of Business, USC, United States
Co-authors: Bikram Karmakar, Ohjin Kwon, Gourab Mukherjee

A spatial autoregressive multinomial probit model is proposed and estimated, in which consumers product preferences are correlated based upon their close they are to each other. The proposed model uses a Bayesian structural uncertainty approach to combine multiple sources of such contiguity information and also incorporates consumer response heterogeneity. The model is applied to the unique problem of improving the efficacy of promotional programs that offer targeted conquesting and loyalty discounts to consumers, which is common in the auto industry but unstudied in the marketing literature. Model calibration on automobile transaction data from the Los Angeles market confirms that previous purchases made by consumers are predictive of the future purchases of other consumers. Targeted discounts derived from the proposed model for conquesting and loyalty promotional programs substantially increase manufacturer profits. We demonstrate that the extant method of using a linear combination of the individual weight matrices provides an inferior fit and lower incremental profits than the proposed Bayesian structural uncertainty approach to information assimilation.

E1211: **Understanding early adoption of hybrid cars via a new multinomial probit model with multiple spatial weights**
Presenter: **Bikram Karmakar**, University of Florida, United States
Co-authors: Ohjin Kwon, Gourab Mukherjee, Sivaramakrishnan Siddarth, Jorge Silva-Risso

A new spatial multinomial probit model is proposed in which the network connectedness of consumers impacts their preference and marketing mix coefficients. Further, these coefficients can be spatially correlated in their unique way. Thus, for example, the utility intercepts may be correlated based on the geographical distance between consumers while the other coefficients may be correlated based on a different contiguity metric built on consumers previous purchase information. We propose a new approach to parameter estimation that significantly expands the scope of our model to handle more consumers and choice alternatives. This method augments the computationally expensive E-step in the Expectation-Maximization algorithm with a fast Gibbs sampling method, divides the M-step into two sub-steps for faster computation, and uses a fast back-fitting method involving a sequence of weighted regressions. We prove the convergence of the algorithm to a local maximum and provide consistent estimators of the standard errors. We use this model on automobile transaction data from the Sacramento market during the first half of 2008. We show how the model helps to gain a better understanding of how consumers adopted hybrid cars during this critical time and demonstrate how an automobile manufacturer can leverage the revealed heterogeneous spatial contiguity effects to develop more effective targeted promotions to accelerate the consumer adoption of a hybrid car.

E1294: **A large-scale mixed joint modeling approach for efficient digital coupon marketing**
Presenter: **Gourab Mukherjee**, University of Southern California, United States

A large-scale joint modeling framework is developed for analyzing customer responses to different types of email coupons. We use finite mixture generalized linear models that capture latent heterogeneous subpopulations in the customer pool and subsequently estimate heterogeneous effects of different coupons after adjusting for customer types, preferences, attributes, and historical behaviors. Analyzing consumer responses to coupons from the apparel industry, we demonstrate the applicability of our method in formulating optimal coupon marketing strategies.

E1507: **Achieving fairness via post-processing in web-scale recommender systems**
Presenter: **Kinjal Basu**, LinkedIn, United States

Building fair recommender systems is a challenging and extremely important area of study due to its immense impact on society. We focus on two commonly accepted notions of fairness for statistical models powering such recommender systems, namely equality of opportunity and equalized odds. These measures of fairness make sure that equally "qualified" (or "unqualified") candidates are treated equally regardless of their protected attribute status (such as gender or race). We propose scalable methods for achieving equality of opportunity and equalized odds in rankings in the presence of position bias, which commonly plagues data generated from recommendation systems. The algorithms are model agnostic in the sense that they depend only on the final scores provided by a model, making them easily applicable to virtually all web-scale recommender systems. We conduct extensive simulations as well as real-world experiments to show the efficacy of our approach.

EO535 Room Virtual R22 BAYESIAN NONPARAMETRICS: MODELING AND COMPUTATION Chair: Federico Camerlenghi

E0390: **Repulsive mixture models: Modelling and computations, with applications to high-dimensional data**
Presenter: **Mario Beraha**, Politecnico di Milano and Universita di Bologna, Italy
Co-authors: Lorenzo Ghilotti, Alessandra Guglielmi

Bayesian mixture models offer a coherent framework for density estimation and model-based clustering. Usual formulations assume that, a priori, the cluster-specific parameters are i.i.d. from some base distribution, which may lead to estimating redundant clusters, especially when the model is misspecified. In repulsive mixtures, a joint prior for cluster-specific parameters is assumed, which puts higher mass on regular point patterns, i.e., well-separated configurations. Building on ideas from normalized random measures, we describe a general framework for repulsive mixture

models, where a random probability measure is derived from a marked Gibbs point process with a (possibly unnormalized) density with respect to the unit rate Poisson point process. In particular, repulsiveness is encouraged among cluster-specific parameters while the unnormalized weights of the mixture arise from independent marks. When considering high-dimensional data, we show that repulsiveness can be incorporated into a latent factor model by means of an anisotropic point process and discuss specifically the case of determinantal point processes. We derive an MCMC sampler to simulate from the posterior distribution and validate our model on synthetic and real data.

E0424: **Gibbs-type random partition aligned on a graph: Application to single-cell RNA data**
Presenter: **Giovanni Rebaudo**, University of Texas at Austin, United States
Co-authors: Peter Mueller

Bayesian nonparametric mixtures and random partition models are effective tools to perform probabilistic clustering. However, standard independent mixture models can be restrictive in some applications, such as inference on cell-lineage due to the biological relations of the clusters. Motivated by single cells RNA application, we develop a novel dependent mixture model to jointly perform cluster analysis and align the cluster on a graph. Our flexible random partition model aligned on a graph cleverly exploits Gibbs type random partition as building blocks allowing suitable variable augmentations to derive analytical results on the partition distribution. From the law of the random partition, we derive a generalisation of the well-known Chinese restaurant process and a related simple MCMC algorithm to perform Bayesian inference. We illustrate the effectiveness of our proposal both on synthetic data and RNA expressions of stem cells.

E0594: **More for less: Predicting and maximizing genetic variant discovery via bayesian nonparametrics**
Presenter: **Lorenzo Masoero**, MIT, United States
Co-authors: Tamara Broderick, Stefano Favaro, Federico Camerlenghi

While the cost of sequencing genomes has decreased dramatically in recent years, this expense often remains non-trivial. Under a fixed budget, then, scientists face a natural trade-off between quantity and quality: spending resources to sequence a greater number of genomes (quantity) or spending resources to sequence genomes with increased accuracy (quality). Our goal is to find the optimal allocation of resources between quantity and quality. Optimizing resource allocation promises to reveal as many new variations in the genome as possible. We introduce a Bayesian nonparametric methodology to predict the number of new variants in a follow-up study based on a pilot study. We validate our method on cancer and human genomics data. When experimental conditions are kept constant between the pilot and follow-up, we find that our prediction is competitive with the best existing methods. Unlike current methods, though, our new method allows practitioners to change experimental conditions between the pilot and the follow-up. We demonstrate how this distinction allows our method to be used for more realistic predictions and for optimal allocation of a fixed budget between quality and quantity.

E0895: **A Bayesian nonparametric approach for inferring drug combination effects on mental health in people with HIV**
Presenter: **Yanxun Xu**, Johns Hopkins University, United States

Although combination antiretroviral therapy (ART) is highly effective in suppressing viral load for people with HIV (PWH), many ART agents may exacerbate adverse effects including depression. Therefore, understanding the effects of ART drugs on mental health can help clinicians personalize medicine with less adverse effects for PWH. The emergence of electronic health records offers researchers unprecedented access to HIV data including individuals' mental health records, drug prescriptions, and clinical information over time. However, modeling such data is very challenging due to the high-dimensionality of the drug combination space, the individual heterogeneity, and the sparseness of the observed drug combinations. We develop a Bayesian nonparametric approach to learn drug combination effect on mental health in PWH adjusting for socio-demographic, behavioral, and clinical factors. The proposed method is built upon the subset-tree kernel method that represents drug combinations in a way that synthesizes known regimen structure into a single mathematical representation. It also utilizes a distance-dependent Chinese restaurant process to cluster heterogeneous populations while taking into account individuals' treatment histories. We apply the method to a dataset from the Women's Interagency HIV Study, yielding interpretable and promising results.

EO360 Room Virtual R23 CAUSAL INFERENCE IN THE ERA OF DATA SCIENCE Chair: Alessandra Mattei

E0474: **Adjusting for nonlocal spatial confounding with U-nets in studies of meteorology and air pollution**
Presenter: **Corwin Zigler**, University of Texas at Austin, United States
Co-authors: Mauricio Tec

Causal effects of spatially-varying exposures on spatially-varying outcomes can be subject to nonlocal confounding, which exists when treatments and outcomes for an index unit are dictated in part by covariates of other (perhaps nearby) units. We offer a deep-learning approach to encode nonlocal covariate information into a vector defined for each observational unit that can be used to adjust for nonlocal confounding. The approach is based on a type of convolutional neural network, called a U-net, that leverages the idea that regional confounding information can be processed in a manner similar to the information contained in an image. We illustrate the approach in two studies of causal effects of air pollution exposure, where meteorology is an inherently regional construct that threatens causal estimates with both local and nonlocal (regional) information. We illustrate the ability of the proposed U-net representation to capture relevant nonlocal confounding information that cannot be fully characterized with simple functions of local and regional meteorological covariates.

E0768: **Relative-risk scale effects in mediation analysis**
Presenter: **Monia Lupparelli**, University of Florence, Italy
Co-authors: Alessandra Mattei

Causal mediation analysis studies the causal pathways of a treatment on an outcome by investigating the role of intermediate variables, named mediators, in the treatment-outcome relationship. When the effect of the treatment on the response might be channelled by mediators, the interest is on disentangling indirect effects, that are through the mediators of interest, and direct effects that are through other pathways other than the mediators. Under a sequential ignorability assumption, we propose a regression-based approach for binary outcomes and multiple mediators so that the total causal relative risk can be decomposed into the natural direct and indirect relative risk by combining model parameters.

E1678: **Estimand strategies for RCTs with intercurrent events**
Presenter: **Fabrizia Mealli**, University of Florence, Italy

The ICH E9(R1) Addendum has provided a framework for discussing interesting estimands that may be of relevance in RCTs affected by the intercurrent events of various types. The Addendum is of fundamental importance because it provides principles regarding the analysis of RCTs and observational studies with intercurrent events. One of the strategies that is proposed is the Principal Stratum strategy. We will discuss issues regarding the application of such a strategy, offering guidelines regarding which data should be collected at baseline and post-treatment to support assumptions and analysis. Various approaches to Principal Stratification analysis will be reviewed with associated carefully planned sensitivity analysis to assess the robustness of conclusions to deviations from such assumptions, as well as to the choice of an analytic approach more broadly (e.g., parametric vs semiparametric or non-parametric). Examples will be used to illustrate concepts.

E1694: **Data-driven heterogeneity detection among subgroup-specific exposure-response functions**
Presenter: **Falco Joannes Bargagli Stoffi**, Harvard University, United States

Over the past several years, various tree-based methods have been developed to identify subgroups of a population with significantly different conditional average treatment effects compared with the population average effect. Despite their success when applied to settings where the

exposure is binary, these methods have not yet been extended to settings with a continuous exposure variable. We develop a flexible method that extends tree-based causal effect heterogeneity identification to settings with continuous exposures, where practitioners can pre-specify a function to summarise exposure-response curves. In the motivating application, we assess the effects of exposure to air pollution (PM2.5) on mortality. While the effect of PM2.5 on mortality has been estimated as an exposure-response curve for various population subgroups, these subgroups are not generally selected for analysis in a data-driven manner. Thus, important subgroups could be overlooked. Our tree-based identification method provides an opportunity to identify important subgroups in a more systematic way.

EO308 Room Virtual R24 HIGH DIMENSIONAL TENSOR REGRESSION Chair: Fei Jiang

E0472: **Covariance estimation for matrix-valued data**
Presenter: **Dehan Kong**, University of Toronto, Canada

Covariance estimation for matrix-valued data has received increasing interest in applications. Unlike previous works that rely heavily on matrix normal distribution assumption and the requirement of fixed matrix size, we propose a class of distribution-free regularized covariance estimation methods for high-dimensional matrix data under a separability condition and a bandable covariance structure. Under these conditions, the original covariance matrix is decomposed into a Kronecker product of two bandable small covariance matrices representing the variability over row and column directions. We formulate a unified framework for estimating bandable covariance and introduce an efficient algorithm based on rank one unconstrained Kronecker product approximation. The convergence rates of the proposed estimators are established, and the derived minimax lower bound shows our proposed estimator is rate-optimal under certain divergence regimes of matrix size. We further introduce a class of robust covariance estimators and provide theoretical guarantees to deal with heavy-tailed data. We demonstrate the superior finite-sample performance of our methods using simulations and real applications from a gridded temperature anomalies dataset and an S&P 500 stock data analysis.

E0481: **High-dimensional tensor autoregression**
Presenter: **Guodong Li**, University of Hong Kong, Hong Kong

Modern technological advances have enabled an unprecedented amount of structured data with complex temporal dependence, urging the need for new methods to efficiently model and forecast high-dimensional tensor-valued time series. Thie aim is to provide the first practical tool to accomplish this task via autoregression (AR). By considering a low-rank Tucker decomposition for the transition tensor, the proposed tensor autoregression can flexibly capture the underlying low-dimensional tensor dynamics, providing both substantial dimension reduction and meaningful dynamic factor interpretation. For this model, we introduce both low-dimensional rank-constrained estimators and high-dimensional regularized estimators. We derive their asymptotic and non-asymptotic properties. In particular, by leveraging the special balanced structure of the AR transition tensor, a novel convex regularization approach, based on the sum of nuclear norms of square matricizations, is proposed to efficiently encourage low-rankness of the coefficient tensor. A truncation method is further introduced to consistently select the Tucker ranks. Simulation experiments and real data analysis demonstrate the advantages of the proposed approach over various competing ones.

E1437: **A unified framework and fast computation for large-margin tensor classifiers**
Presenter: **Boxiang Wang**, University of Iowa, United States
Co-authors: Qing Mai

Tensor data, also known as higher-order arrays, are increasingly common in econometrics, image processing, social network analysis, digital marketing, among many other applications. We focus on binary classification and we formulate a unified framework for tensor large-margin classifiers. The framework includes some popular classifiers such as support vector machine (SVM), Huberized SVM, distance-weighted discrimination, and logistic regression. Despite the success of these classifiers in classifying the vector-valued data, the computation is actually highly intensive. Although it seems natural to extend these methods to the tensor data analysis by applying an alternating minimization-type algorithm, this approach is rather computationally prohibitive. To over such computational burdens, we develop a computationally efficient accelerated proximal gradient descent algorithm to solve the smooth tensor large-margin classifiers and show the convergence. We also develop a novel smoothing algorithm to solve the tensor SVM. In addition, we reveal some connections between our unified framework and the tensor single index model. We use simulations and real applications to demonstrate the performance and efficiency of our proposal.

E0417: **Bayesian inference for matrix-valued data**
Presenter: **Weining Shen**, UC Irvine, United States

A Bayesian nonparametric matrix clustering approach is proposed to analyze the latent heterogeneity structure in the shot selection data collected from professional basketball players in the National Basketball Association (NBA). The proposed method adopts a mixture of finite mixtures framework and fully utilizes the spatial information via a mixture of matrix normal distribution representation. We propose an efficient Markov chain Monte Carlo algorithm for posterior sampling that allows simultaneous inference on both the number of clusters and the cluster configurations. We also establish large-sample convergence properties for the posterior distribution. The compelling empirical performance of the proposed method is demonstrated via simulation studies and an application to shot chart data from selected players in the NBAs 2017-2018 regular season.

EO102 Room Virtual R25 COPULAS AND DEPENDENCE MODELLING II Chair: Piotr Jaworski

E0904: **On attainability of Kendall's tau matrices and concordance signatures**
Presenter: **Johanna Neslehova**, McGill University, Canada
Co-authors: Alexander Alexander John McNeil, Andrew Smith

The concordance signature of a random vector or its distribution is defined to be the set of concordance probabilities for margins of all orders. We will show that the concordance signature of a copula is always equal to the concordance signature of some unique mixture of so-called extremal copulas. This result has a number of interesting consequences, which we will explore, such as a characterization of the set of Kendall rank correlation matrices as the cut polytope, and a method for determining whether a set of concordance probabilities is attainable. We also use it to show that the widely-used elliptical distributions yield a strict subset of the attainable concordance signatures as well as a strict subset of the attainable Kendall rank correlation matrices, and prove that the Student t copula converges to a mixture of extremal copulas sharing its concordance signature with all elliptical distributions that have the same correlation matrix. Finally, we will discuss a method of estimating an attainable concordance signature from data, and highlight applications to Monte Carlo simulations of dependent random variables as well as expert elicitation of consistent systems of Kendall's tau dependence measures.

E1047: **Conditional copulas: Mean, median and quantiles**
Presenter: **Irene Gijbels**, KU Leuven, Belgium
Co-authors: Margot Matterne

A conditional copula fully characterizes the dependency between random variables, conditionally upon a covariate (vector). A way to summarize this dependence structure taking into account the impact of the covariate is via the conditional copula, which under fairly general conditions coincides with the partial copula. A mean is just one way to summarize this conditional dependence behavior. We introduce the notions of median conditional copula, and more generally quantile conditional copula. We investigate the existence of these concepts, and establish explicit expressions for calculating them. Examples are given to illustrate the concepts.

E1038: **Algorithms for constructing copulas via *-product decompositions**
Presenter: **Enrique de Amo**, University of Almeria, Spain

Although the theoretical problem for constructing a copula for two given measure-preserving functions is completely solved, for practical purposes it is a rather difficult task. We provide explicit algorithms which solve this problem in various contexts such as the measure-preserving functions are monotonic, as well as the copula is an extreme copula, a diagonal copula, an extremal biconic copula, an Archimedean copula, a conic copula, or a migrative copula.

E1139: **On left truncation invariant limit distributions under low threshold**
Presenter: **Claudio Ignazzi**, Universita del Salento, Lecce, Italy, Italy
Co-authors: Fabrizio Durante, Piotr Jaworski

Given a random pair (X,Y) distributed according to $C(F,F)$, where C is a bivariate copula and F is a continuous univariate distribution function, the limit distribution of (X,Y) given that the values of X fall under a low threshold is studied. This limit distribution is defined in three different ways depending on which of the three classes of univariate marginals includes F, based on its rate of decay at minus infinity: Frechet, Weibull or Gumbel. Various assumptions on the copula C are needed, such as exchangeability as well as information on its tail behavior, such as having a non-zero lower tail dependence coefficient. After computing the above limit distributions in the three cases considered, the (unique) copula of the three limit distributions is determined. It turns out to be the same copula for all three cases. Finally, the above copula is proved to be invariant under univariate truncation of the first variable, or left truncation invariant.

EO426 Room Virtual R26 STATISTICAL ASPECTS OF MEASUREMENT AND PSYCHOMETRICS Chair: Daphna Harel

E0536: **Using optimal test assembly to shorten patient reported outcome measures**
Presenter: **Daphna Harel**, New York University, United States

Patient-reported outcome measures are widely used to assess respondent experiences, well-being, and treatment response in clinical trials and cohort-based observational studies in both medicine and psychological studies. However, respondents may be asked to respond to many different scales in order to provide researchers and clinicians with a wide array of information regarding their experiences. Therefore, collecting such long and cumbersome patient-reported outcome measures may burden respondents and increase research costs. However, little research has been conducted on optimal, replicable, and reproducible methods to shorten these instruments. We propose the use of mixed-integer programming through Optimal Test Assembly as a method to shorten patient-reported outcome measures. We will provide several examples as well as a comparison to existing methods in the field.

E0605: **On the detection of bots in online surveys**
Presenter: **Carl Falk**, McGill University, Canada
Co-authors: Michael John Ilagan

Academic research via online data collection of survey responses through crowdsourcing platforms has become increasingly prevalent in the social sciences. However, the increased anonymity of participants coupled with monetary compensation may result in the contamination of such data by bots or random responders. While a number of outlier detection indices are often recommended to detect bots, their practical combined usage is hampered by the lack of recommendations for empirically derived cut-off values. We propose and compare four algorithms that could be used to classify bots in an unsupervised manner while leveraging such outlier detection indices. The basis of these algorithms relies on the assumptions that bots are exchangeable random vectors and that detection indices tend to separate humans and bots. Permutations are then used to derive an accompanying test and/or inform clustering techniques. In simulations, some studied techniques achieved about 90-95% accuracy across conditions ranging from low bot contamination (5%) to high bot contamination (95%). Given that data collection often occurs in the presence of multi-item scales with Likert-type items, additional discussion focuses on 1) scale/study design conditions under which we would expect such algorithms to encounter difficulty; and 2) the potential for indices derived from psychometric models (e.g., based on item response theory) to be able to detect clusters of bots.

E0691: **Applications of Bayesian item response theory**
Presenter: **Chelsea Parlett**, Chapman University, United States
Co-authors: Erik Linstead, Susanne Jaeggi, Grace Lin

Item Response Theory (IRT) models are common in psychometrics, but can have applications to many other fields. The nature of IRT leads to a large number of model parameters (at minimum one parameter per subject and one per item), which can be difficult to estimate in smaller datasets where IRT may otherwise be useful. In addition to the typical benefits of Bayesian models, additional information and/or precision from priors can be helpful in fitting IRT models with many parameters. We will explore Bayesian IRT models using the language Stan, and extend the IRT framework to Beta Regression using Stan. Applications to behavioral data will also be shown.

E0736: **Using computer adaptive testing to improve migraine outcomes**
Presenter: **Erin Buchanan**, Harrisburg University of Science and Technology, United States

Computer adaptive testing allows for personalized examination of an underlying trait to precisely pinpoint an individuals' measurement of their trait level. Migraine is a multifaceted disease, with varied symptomology and treatment options. Current assessment tools for migraine focus on head pain, ignoring patient concerns beyond pain relief such as social interaction, completing day-to-day activities, and provider and financial worries. Recent funding opportunities indicate an interest in providing digital healthcare solutions that are tailored to patients for improved shared decision making at the point of care. Statistical tools can provide the necessary customization of measurement delivery, and this presentation will focus on the application of item response theory and computer adaptive testing to the assessment of migraine symptomology. The item response theory and adaptive testing approach will be contrasted with the traditional development of measurement scales using classical test theory and exploratory factor analysis to demonstrate statistical results that each provides for patient reported-outcome measurement.

EO649 Room Virtual R27 STATISTICAL LEARNING AND INFERENCE ON COMPLEX DATA STRUCTURES Chair: Tianxi Li

E0867: **Spectral analysis of networks with latent space dynamics and signs**
Presenter: **Joshua Cape**, University of Pittsburgh, United States

The problem of modeling and analyzing latent space dynamics in collections of networks is considered. Towards this end, we pose and study latent space generative models for signed networks that are amenable to inference via spectral methods. Permitting signs, rather than restricting to unsigned networks, enables richer latent space structure and permissible dynamic mechanisms that can be provably inferred via low-rank truncations of observed adjacency matrices. The treatment of and ability to recover latent space dynamics holds across different levels of granularity, namely, at the overall graph level, for communities of nodes, and even at the individual node level. We provide synthetic and real data examples to illustrate the effectiveness of methodologies and to corroborate the accompanying theory. The contributions complement an emerging statistical paradigm for random graph inference encompassing random dot product graphs and generalizations thereof.

E1145: **Modeling continuous-time networks of relational events**
Presenter: **Subhadeep Paul**, The Ohio State University, United States

Spatiotemporal data with complex network dependencies are increasingly available in many application problems involving human mobility, geo-

tagged social media, disease transmission, international relationships and conflict. In many such application settings involving spatiotemporal data, the observed data consist of timestamped relational events. For example, in online social media, users interact with each other through events that occur at specific time instances such as liking, mentioning, commenting, or sharing another user's content. In international relations and conflicts, nations commit acts of hostility or disputes through discrete time-stamped events. We will introduce statistical models and methods for analyzing such datasets combining tools from network analysis and multivariate point processes. We will also describe scalable estimation methods and study the asymptotic properties of the estimators. Finally, we will demonstrate the models are able to fit several real datasets well and predict temporal structures in those datasets.

E1205: **Using maximum entry-wise deviation to test the goodness-of-fit for stochastic block models**
Presenter: **Emma Jingfei Zhang**, University of Miami, United States

The stochastic block model is widely used for detecting community structures in network data. How to test the goodness-of-fit of the model is one of the fundamental problems and has gained growing interest in recent years. We propose a novel goodness-of-fit test based on the maximum entry of the centered and re-scaled adjacency matrix for the stochastic block model. One noticeable advantage of the proposed test is that the number of communities can be allowed to grow linearly with the number of nodes ignoring a logarithmic factor. We prove that the null distribution of the test statistic converges in distribution to a Gumbel distribution, and we show that both the number of communities and the membership vector can be tested via the proposed method. Further, we show that the proposed test has an asymptotic power guarantee against a class of alternatives. We also demonstrate that the proposed method can be extended to the degree-corrected stochastic block model. Both simulation studies and real-world data examples indicate that the proposed method works well.

E1388: **Reluctant interaction modeling in GLMs**
Presenter: **Guo Yu**, University of California Santa Barbara, United States

While including pairwise interactions in a regression model can better approximate the response surface, fitting such an interaction model is a well-known difficult problem. In particular, analyzing contemporary high-dimensional datasets often leads to extremely large-scale interaction modeling problem, where the challenge is posed to identify important interactions among millions or even billions of candidate interactions. While several methods have recently been proposed to tackle this challenge, they are mostly designed by (1) focusing on linear models with interactions and (or) (2) assuming the hierarchy assumption among the important interactions. In practice, however, neither of these two building blocks has to hold. We propose an interaction modeling framework in generalized linear models (GLMs) which is free of any assumptions on hierarchy. The basic premise is a non-trivial extension of the reluctance principle to interaction selection in GLMs, where main effects are preferred over interactions if all else is equal. The proposed method is easy to implement, and is highly scalable to large-scale datasets. We show favorable theoretical properties of the proposed method. Numerical results show that the proposed method does not sacrifice any statistical performance in the presence of significant computational gain.

EO659 Room Virtual R28 BAYESIAN METHODS IN CAUSAL INFERENCE Chair: Chanmin Kim

E0440: **Bayesian machine learning for causal inference with multiple treatments and multilevel censored survival outcomes**
Presenter: **Liangyuan Hu**, Rutgers University, United States
Co-authors: Jiayi Ji, Joseph Hogan

Despite numerous recent advances in causal inference, the literature for handling data with multiple treatments and multilevel censored survival outcomes is sparse. Here we develop a way to use Bayesian Additive Regression Trees, a likelihood-based machine learning modeling technique, to draw causal inferences about the effects of multiple treatments for clustered observational survival data. This approach will provide substantial modeling flexibility for a data structure for which few off-the-shelf causal inference methods are available. We further develop a flexible and interpretable sensitivity analysis framework to handle the no unmeasured confounding assumption, respecting the multilevel survival data structure. Our approach addresses unmeasured confounding at both cluster and individual levels and incorporates uncertainty about unidentified model components formally into the analysis. The operating characteristics of our proposed method are examined via an extensive simulation. We demonstrate the developed methods via a case study evaluating the survival effects of three popular types of treatments for high risk localized prostate cancer using the national cancer database.

E0596: **The impact of positivity assumption on causal inference using Bayesian nonparametric methods**
Presenter: **Jason Roy**, Rutgers University, United States
Co-authors: Nandita Mitra, Yaqian Zhu

In observational studies, differences between the treatment and control groups may be due to confounding variables. To assess the causal effects of a treatment in a population, an important identifiability condition is the positivity assumption (or 'overlap'), which requires the probability of treatment to be bounded away from 0 and 1. That is, for every covariate combination, we should be able to observe both treatment and control subjects if the sample size is large enough. We discuss how different causal inference methods (parametric and Bayesian non-parametric) implicitly deal with non-overlap. We assess the performance of these approaches with respect to bias and efficiency in simulations.

E0639: **A nonparametric Bayesian model-based construction of synthetic treatment arms in a clinical study**
Presenter: **Peter Mueller**, UT Austin, United States

Randomized clinical trials (RCT) are the gold standard for approvals by regulatory agencies. However, RCT's are increasingly time-consuming, expensive, and laborious with a multitude of bottlenecks involving volunteer recruitment, patient truancy, and adverse events. An alternative that fast tracks clinical trials without compromising the quality of scientific results is desirable to more rapidly bring therapies to consumers. We propose a model-based approach using nonparametric Bayesian common atoms models for patient baseline covariates. This specific class of models has two critical advantages in this context: (i) The models have full prior support, i.e., allow to approximate arbitrary distributions without unreasonable restrictions or shrinkage in specific parametric families; (ii) inference naturally facilitates a reweighting scheme to achieve equivalent populations. We prove the equivalence of the synthetic and other patient cohorts using an independent separate verification. Failure to classify a merged data set using a flexible statistical learning method such as random forests, support vector machines etc. proves equivalence. We implement the proposed approach in two motivating case studies.

E1052: **Imputing biomarkers from cognitive assessments: combating covariate shift by assuming causal stationarity**
Presenter: **Chelsea Krantsevich**, Arizona State University, United States
Co-authors: Richard Hahn

Motivated by the problem of developing accurate biomarkers to track the progression of Alzheimer's disease, the aim is to consider how incorporating a causal understanding of the underlying biology can improve the prediction of biomarker trajectories. We introduce a causal imputation method based on biologically-motivated causal graphs and compare its performance to an unconstrained supervised learning method that ignores causal relationships. We demonstrate that the causal approach is substantially more accurate in the presence of "covariate shift", where the test population differs in important but unforeseen ways from the training population.

EO424 Room Virtual R29 ADVANCES ON BAYESIAN COMPUTATION AND ITS APLLICATIONS Chair: Brenda Betancourt

E0382: **Stratified stochastic variational inference for network factor models**
Presenter: **Emanuele Aliverti**, University Ca' Foscari of Venezia, Italy

Recently, there has been considerable interest in the Bayesian modeling of networks using latent space models. As the number of nodes increases, Markov Chain Monte Carlo can be demanding, thus motivating research into alternative algorithms that scale well in high dimensions. The focus is on the latent factor model for networks and on scalable algorithms to perform approximate Bayesian inference. Leveraging sparse representations of network data and conditionally conjugate specifications, a stratified stochastic variational algorithm is developed. Empirical results demonstrate the benefit of the proposed specification in terms of computational resources and timing.

E0553: **Sampling from multimodal target distributions using tempered Hamiltonian transitions**
Presenter: **Joonha Park**, University of Kansas, United States

Hamiltonian Monte Carlo (HMC) methods are widely used to draw samples from unnormalized target densities due to high efficiency and favorable scalability with respect to increasing space dimensions. However, HMC struggles when the target distribution is multimodal, because the maximum increase in the potential energy function (i.e., the negative log density function) along the simulated path is bounded by the initial kinetic energy, which follows a half of the chi-squared distribution with d degrees of freedom, where d is the space dimension. We develop a Hamiltonian Monte Carlo method that can construct paths that cross high potential energy barriers. This approach does not require the modes of the target distribution to be known. Our method constructs the Hamiltonian paths while continuously increasing and decreasing the mass of the simulated particle, and thus it can be viewed as a case of the tempered transitions method. We develop a practical tuning strategy for the mass schedule, aiming to achieve high mode-hopping frequency. In addition to highly competitive scalability with dimensions, our method has a practical advantage over other tempering methods in the Gibbs sampler settings, where the target distribution changes frequently. We demonstrate that our method can facilitate frequent mode hopping in high-dimensional distributions using mixtures of normal distributions and a sensor network self-localization problem.

E0699: **Divide-and-conquer Bayesian inference in hidden Markov models**
Presenter: **Sanvesh Srivastava**, The University of Iowa, United States

The focus is on divide-and-conquer Bayesian inference in models for dependent data. Divide-and-conquer Bayesian methods consist of three steps: dividing the data into smaller computationally manageable subsets, running a sampling algorithm parallel on all the subsets, and combining parameter draws from all the subsets. The combined parameter draws are used for efficient posterior inference in massive data settings. Several innovative methods have been developed over the years, but a major restriction common to all is that their first two steps assume that the observations are independent. We address this problem by developing a divide-and-conquer method for Bayesian inference in parametric hidden Markov models, where the state space is known and finite. First, we show that the dependence can be preserved on the subsets by appropriately modifying the subset likelihoods. Second, if the number of subsets is chosen appropriately depending on the mixing properties of the hidden Markov chain, then we show that the subset posterior distributions defined using the modified likelihood are asymptotically normal as the subset sample size tends to infinity. Finally, we present numerical results to justify the empirical validity of the theoretical results.

E1380: **A Bayesian approach to streaming multi-file record linkage**
Presenter: **Ian Taylor**, Colorado State University, United States
Co-authors: Andee Kaplan, Brenda Betancourt

Record linkage is the task of combining records from multiple files which refer to overlapping sets of entities when there is no unique identifying field in the records. In streaming record linkage, files arrive in time and estimates of links are desired after the arrival of each file. This problem arises in settings such as longitudinal surveys. The challenge in streaming record linkage is efficiently updating parameter estimates as new files arrive. We approach the problem from a Bayesian perspective with estimates in the form of posterior samples of parameters and present a method for updating link estimates after the arrival of a new file that is faster than starting an MCMC from scratch. We generalize a Bayesian Fellegi-Sunter model for two files and apply Sequential Markov Chain Monte Carlo for streaming sample updates. We examine the effect of the prior distribution and the strength of the prior information on the resulting estimates. We apply this method to simulated data and data from the Social Diagnosis Survey of Polish households.

EO571 Room Virtual R30 APPLIED STATISTICAL LEARNING Chair: Alejandro Murua

E0560: **Making probabilistic predictions in multi-response regression problems**
Presenter: **Mu Zhu**, University of Waterloo, Canada
Co-authors: Marius Hofert, Avinash Prasad

A fully nonparametric approach to probabilistic predictions in multi-response regression problems is proposed. The main focus is on learning the dependence between multiple responses with a generative neural network, and demonstrating through a variety of data sets that the flexibility afforded by being fully nonparametric does make a difference.

E0690: **A twin neural model for causal inference: Applications in python**
Presenter: **Mouloud Belbahri**, University of Montreal, Canada
Co-authors: Olivier Gandouet

The focus is on the prediction of heterogeneous treatment effects in the randomized case of causal inference. We developed a solution for a specific twin neural network architecture allowing joint optimization of counterfactual marginal probabilities. We show that this model is a generalization of the logistic interaction model. We train our models with a new loss function, defined by taking advantage of a link with the Bayesian interpretation of relative risk. We modify the stochastic gradient descent algorithm to allow sparse structured solutions. This helps training to a great extent. We show that our method is competitive with the state of the art on real data from large-scale marketing campaigns.

E1385: **Estimating the number of components in finite mixture models via the group-sort-fuse procedure**
Presenter: **Abbas Khalili**, McGill University, Canada
Co-authors: Tudor Manole

Estimation of the number of components (or order) of a finite mixture model is a long-standing and challenging problem in statistics. We propose the Group-Sort-Fuse (GSF) procedure, a new penalized likelihood approach for simultaneous estimation of the order and mixing measure in multidimensional finite mixture models. Unlike methods that fit and compare mixtures with varying orders using criteria involving model complexity, our approach directly penalizes a continuous function of the model parameters. More specifically, given a conservative upper bound on the order, the GSF groups and sorts mixture component parameters to fuse those which are redundant. For a wide range of finite mixture models, we show that the GSF is consistent in estimating the true mixture order. The GSF is implemented for several univariate and multivariate mixture models in the R package GroupSortFuse. Its finite sample performance is supported by a thorough simulation study, and its application is illustrated on two real data examples.

E1460: **Data challenges in applied AI projects: A few stories**
Presenter: **Jean-Francois Plante**, HEC Montreal, Canada

Machine learning algorithms are designed to leverage structure in data in order to discover patterns, and/or to make predictions. In applied projects,

the success of good ideas depends heavily on the availability of appropriate data. The crucial role of those data is typically not emphasized enough, and industrial partners may be sceptical about it. We will present a few recent examples of industrial projects where good ideas were limited by the availability or characteristics of the data. Additional challenges also come when partners get excited by generic promises that may not be applicable to their project. From sawmills to ventilation systems, we will share some key steps of three different industrial research projects in applied AI.

EO724 Room Virtual R31 NEW APPLICATIONS AND DIRECTIONS IN STATE SPACE MODELING Chair: Daniel McDonald

E1499: **State-space models in ecology: Opportunities and challenges**
Presenter: **Marie Auger-Methe**, The University of British Columbia, Canada

State-space models (SSMs) are increasingly used in ecology to model time-series such as animal movement paths and population dynamics. This type of hierarchical model is structured to account for two levels of variability: biological stochasticity and measurement error. Because they can account for large measurement error, they are particularly popular to study marine animals for which it is often hard to get accurate time-series of geographic locations and population counts. SSMs are flexible. They can model linear and nonlinear processes using a variety of statistical distributions. We will use marine movement data to introduce SSMs and to demonstrate when these models are useful and when they can fail. We will also highlight new tools that can help fit state-space models to data.

E1527: **Markov-switching state-space models with applications to neuroimaging**
Presenter: **David Degras**, University of Massachusetts Boston, United States
Co-authors: Hernando Ombao, Chee Ming Ting

State-space models (SSM) with Markov switching offer a powerful framework for detecting multiple regimes in time series, analyzing mutual dependence and dynamics within regimes, and assessing transitions between regimes. These models however present considerable computational challenges due to the exponential number of possible regime sequences to account for. In addition, the high dimensionality of time series can hinder likelihood-based inference. To address these challenges, novel statistical methods for Markov-switching SSMs are proposed using maximum likelihood estimation, Expectation-Maximization (EM), and parametric bootstrap. Solutions are developed for initializing the EM algorithm, accelerating convergence, and conducting inference. These methods, which are ideally suited to massive spatio-temporal data such as brain signals, are evaluated in simulations and applications to EEG studies of epilepsy and motor imagery are presented.

E0551: **Markov-switching State Space models for uncovering musical interpretation**
Presenter: **Daniel McDonald**, University of British Columbia, Canada

For concertgoers, musical interpretation is the most important factor in determining whether or not we enjoy a classical performance. Every performance includes mistakes — intonation issues, a lost note, an unpleasant sound — but these are all easily forgotten (or unnoticed) when a performer engages her audience, imbuing a piece with novel emotional content beyond the vague instructions inscribed on the printed page. While music teachers use imagery or heuristic guidelines to motivate interpretive decisions, combining these vague instructions to create a convincing performance remains the domain of the performer, subject to the whims of the moment, technical fluency, and taste. We use data from the CHARM Mazurka Project — forty-six professional recordings of Chopin's Mazurka Op. 63 No. 3 by consummate artists — with the goal of elucidating musically interpretable performance decisions. Using information on the inter-onset intervals of the note attacks in the recordings, we apply functional data analysis techniques enriched with prior information gained from music theory to discover relevant features and perform hierarchical clustering. The resulting clusters suggest methods for informing music instruction, discovering listening preferences, and analyzing performances.

E1703: **State-space epidemiological compartmental models: Approximations, control, forecasting, and real-world complications**
Presenter: **Logan Brooks**, Carnegie Mellon University, United States

Epidemiological compartmental models describe the state of a population using the amounts of individuals falling into certain predefined categories encapsulating their health status, location, age, and/or other characteristics. The evolution of the population state describes the transition of individuals between compartments as their status changes due to interactions with other individuals or the passage of time. The mean change in a compartment's membership at some time is typically a quadratic function of the population's current state. These dynamics can be difficult to analyze, and paired with an observation model, challenging to fit. Linear dynamical approximations lead to statements about herd immunity and objectives for disease control. Similarly, various manipulations of the state-space model equations motivate additional types of approximations and compartmental-model-inspired forecasting frameworks; sufficiently constrained models can also be fitted with general-purpose frameworks. However, real-world details of epidemiological surveillance systems and disease dynamics create complications for common model structures. We will discuss some of these compartmental modeling frameworks, approximations, and complications in the context of seasonal epidemics of influenza-like illness and the COVID-19 pandemic.

EO497 Room Virtual R35 ADVANCES IN STATISTICAL LEARNING AND INFERENCE WITH ROBUST INSIGHTS Chair: Zhao Ren

E0378: **Uncertainty quantification in the Bradley-Terry-Luce model**
Presenter: **Anderson Ye Zhang**, University of Pennsylvania, United States

Ranking from pairwise comparisons is a central problem in a wide range of learning and social contexts. The Bradley-Terry-Luce (BTL) model is one of the most studied models for analyzing ranking data. Despite all the recent progress, uncertainty quantification under the BTL model remains unclear. To address this challenge, we first establish non-asymptotic entrywise distributions of the maximum likelihood estimation and the spectral method under the BTL model. We then develop statistical inference procedures for individual rankings and preference parameters.

E1278: **Large-scale inference of multivariate regression for heavy-tailed and asymmetric data**
Presenter: **Wen Zhou**, Colorado State University, United States
Co-authors: Youngseok Song, Wenxin Zhou

Large-scale multivariate regression is a fundamental statistical tool that finds applications in a wide range of areas. The focus is on the problem of simultaneously testing a large number of general linear hypotheses, encompassing covariate-effect analysis, analysis of variance, and model comparisons. The new challenge that comes along with the overwhelmingly large number of tests is the ubiquitous presence of heavy-tailed and/or highly skewed measurement noise, which is the main reason for the failure of conventional least-squares based methods. For large-scale multivariate regression, we develop a set of robust inference methods to explore data features, such as heavy tailedness and skewness, which are invisible to the scope of least squares. The new testing procedure is built on data-adaptive Huber regression, and a new covariance estimator of regression estimates. Under mild conditions, we show that our methods produce consistent estimates of the false discovery proportion. Extensive numerical experiments, along with an empirical study on quantitative linguistics, demonstrate the ad- vantage of our proposal compared to many state-of-the-art methods when the data are generated from heavy-tailed and/or skewed distributions.

E1300: **Conquer: Convolution-smoothed quantile regression**
Presenter: **Wenxin Zhou**, University of California San Diego, United States
Co-authors: Kean Ming Tan, Lan Wang

Quantile regression is a powerful tool for learning the relationship between a response variable and a multivariate predictor while exploring heterogeneous effects. We discuss a convolution-smoothed approach for quantile regression, which is particularly suited for large-scale problems

in both "increasing dimension" and "high-dimensional" regimes. This method, which we refer to as conquer, turns the non-differentiable check loss into a twice-differentiable, convex, and locally strongly convex surrogate, and therefore admits fast and scalable gradient-based algorithms to perform optimization.

E1302: **Retire: Robustified expectile regression in high dimensions**
Presenter: **Kean Ming Tan**, University of Michigan, United States
Co-authors: Rebeka Man, Zian Wang, Wenxin Zhou

High-dimensional data can often display heterogeneity due to heteroscedastic variance or inhomogeneous covariate effects. Penalized quantile and expectile regression methods offer useful tools to detect heteroscedasticity in high-dimensional data. The former is computationally challenging due to the non-smooth nature of the check loss, and the latter is sensitive to heavy-tailed error distributions. We propose and study penalized robustified expectile regression (retire) in high dimensions, with a focus on concave regularization which reduces the estimation bias from l_1-penalization and leads to oracle properties. Theoretically, we establish the statistical properties of the solution path of iteratively reweighted l_1-penalized retire estimation, adapted from the local linear approximation algorithm for folded concave regularization. Under a mild minimum signal strength condition, we show that after as many as $\log\log(d)$ iterations the final estimator enjoys the oracle convergence rate. At each iteration, the weighted l_1-penalized convex program can be efficiently solved by a semismooth Newton coordinate descent algorithm. Numerical studies demonstrate the competitive performance of the proposed procedure compared with either non-robust or quantile regression-based alternatives.

EO088 Room Virtual R36 CAUSAL INFERENCE IN THE PRESENCE OF COMPETING EVENTS Chair: Daniel Nevo

E0430: **Estimating gestational age-specific exposure effects during pregnancy with observational data: A target trials approach**
Presenter: **Mireille Schnitzer**, Universite de Montreal, Canada
Co-authors: Steve Ferreira Guerra, Cristina Longo, Lucie Blais, Robert Platt

Many studies seek to evaluate the effects of potentially harmful pregnancy exposures during specific gestational periods. We consider an observational pregnancy cohort where women can initiate medication usage or become exposed to a drug at various times during their pregnancy. An important statistical challenge involves defining and estimating exposure effects when pregnancy loss or delivery can occur over time. Without proper consideration, the results of standard analyses may be vulnerable to selection bias, immortal time-bias, and time-dependent confounding. We apply the target trials framework of Hernan and Robins in order to define effects based on the counterfactual approach often used in causal inference. This effect is defined relative to a hypothetical randomized trial of timed pregnancy exposures where delivery may precede (and thus interrupt) exposure initiation. We demonstrate tailored implementations of inverse probability weighting (IPW), G-Computation, and Targeted Maximum Likelihood Estimation (TMLE) to estimate the effects of interest. We then apply our proposed methods to a pharmacoepidemiology study to evaluate the potentially time-dependent effect of exposure to inhaled corticosteroids on birth weight in pregnant women with mild asthma.

E0531: **Separable direct and indirect effects in a competing risk setting**
Presenter: **Torben Martinussen**, University of Copenhagen, Denmark

Many research questions involve time-to-event outcomes that can be prevented from occurring due to competing events. In these settings, we must be careful about the causal interpretation of classical statistical estimands. In particular, estimands on the hazard scale, such as ratios of cause-specific or subdistribution hazards, are fundamentally hard to interpret causally. Estimands on the risk scale, such as contrasts of cumulative incidence functions, do have a clear causal interpretation, but they only capture the total effect of the treatment on the event of interest; that is, effects both through and outside of the competing event. To disentangle causal treatment effects on the event of interest and competing events, the separable direct and indirect effects were recently introduced. Here we provide new results on the estimation of direct and indirect separable effects in continuous time. In particular, we derive the nonparametric influence function in continuous time and use it to construct an estimator that has certain robustness properties. We also propose a simple estimator based on semiparametric models for the two cause-specific hazard functions. We describe the asymptotic properties of these estimators. Finally, we suggest extensions to so-called semi-competing events.

E0915: **The subtype-free average causal effect for disease heterogeneity studies**
Presenter: **Daniel Nevo**, Tel Aviv University, Israel
Co-authors: Amit Sasson

A common goal in molecular pathological epidemiology studies is to evaluate whether the effects of risk factors on disease incidence vary across different disease subtypes. A popular approach implements a multinomial regression in which each of the non-zero values corresponds to a bona fide disease subtype. Then, heterogeneity in the exposure effects across subtypes is examined by comparing the coefficients of the exposure between the different subtypes. We explain why this common approach does not recover causal effects, even when all confounders are measured, due to a built-in selection bias in the multinomial regression model. We further develop the Subtype-Free Average Causal Effect (SF-ACE), a well-defined causal effect inspired by the Survivor Average Causal Effect (SACE). We propose identification and estimation approaches for the SF-ACE under different sets of assumptions. Similar to the SACE, the assumptions underlying the identification of the SF-ACE from the data are untestable and can be too strong in some scenarios. Therefore, we also develop a sensitivity analysis to relax some of these assumptions. Finally, we apply our methodology to data from two large cohort studies to study the heterogeneity in the causal effect of smoking on colorectal cancer subtyped by microsatellite status.

E1232: **Causal inference when resources are constrained**
Presenter: **Mats Stensrud**, Ecole polytechnique federale de Lausanne, Switzerland

Many treatments are of limited supply and cannot be provided to all individuals in need. For example, many hospitals have experienced shortages of ventilators, protective equipment and personnel during the COVID-19 pandemic. Similarly, there are waiting lists for organ transplants in most health care systems because suitable organs constitute a limited resource. When resources are limited, policymakers often raise questions about the effects of allocation strategies where only a limited proportion of individuals receive treatment at a given time. We will present current work on causal estimands that target such questions. We will emphasize why existing causal inference methods cannot be used, and we will also give new identifiability results and suggest new estimators.

EO846 Room Virtual R37 TIME SPACE MODELS: EVENTS AT RANDOM BEYOND GAUSSIANITY I Chair: Krzysztof Podgorski

E0846: **Volatility leverage and non-Gaussian shocks**
Presenter: **Farrukh Javed**, Orebro University, Sweden
Co-authors: Krzysztof Podgorski

A general framework in which the leverage effects can be accounted for in volatility models featuring non-Gaussian shocks is introduced and discussed. Some specifications within this framework have appeared in the literature but the natural general structure and its consequences for the meaningful modeling of the leverage effect has not been explained. The framework allows treating in a unified way a large number of non-Gaussian innovation alternatives for modeling financial volatility. The stationarity conditions, moments, dependence structure to account for heavy tails and leverage in the data are discussed. Finally, through empirical investigation, the model efficiency has been evaluated using some benchmark financial data.

E1535: **Signals featuring harmonics with random frequencies: Spectral, distributional and ergodic properties**
Presenter: **Anastassia Baxevani**, University of Cyprus, Cyprus
Co-authors: Krzysztof Podgorski
An interesting class of non-Gaussian stationary processes is obtained when in the harmonics of a signal with random amplitudes and phases, one allows also for frequencies to vary randomly. In the resulting models, the statistical distribution of frequencies determines the process spectrum while the distribution of amplitudes governs the process distributional properties. Since decoupling the distribution from the spectrum can be advantageous in applications, we thoroughly investigate a variety of properties exhibited by these models. A process in the considered class of models is uniquely defined by a triple consisting of a positive scale, a normalized spectrum (which is also the distribution of the frequencies), and a normalized Levy measure determining the process distribution. We extend previous work that represented processes as a finite sum of harmonics, by conveniently embedding them into the class of harmonizable processes. Harmonics are integrated with respect to independently scattered second-order non-Gaussian random measures. We present a proper mathematical framework that allows for studying spectral, distributional, and ergodic properties.

E1564: **Estimation of parameters of elliptic SPDE models driven by non-Gaussian white noise**
Presenter: **Alexandre de Bustamante Simas**, King Abdullah University of Science and Technology, Saudi Arabia
Co-authors: David Bolin, Jonas Wallin
A class of linear elliptic SPDE models driven by type-G Levy Noise will be briefly presented. We will discuss the estimation of parameters in such models and present a stochastic gradient descent approach to estimation. This estimation relies on a Gibbs sampler algorithm, which we show is geometrically ergodic and discuss how the parameters affect its convergence to equilibrium.

E1740: **Multivariate model of precipitation events in California: The role of atmospheric rivers, topography, and climate change**
Presenter: **Anna Panorska**, University of Nevada, United States
Co-authors: Francesco Zuniga, Alexander Weyant, Ilaria Vinci, Alexander Gershunov, Tomasz Kozubowski
A new modeling approach is presenting which describes the joint behavior of the duration, the magnitude, and the peak value of precipitation events in California, USA. The study shows that the model parameters respond to the expected changes in the characteristics of the events such as their origin or the local topography. We also show the response of the parameters to the observed and projected climatic changes during the time period from 1950 to 2100. The joint information on the properties of events was not previously available as the majority of the modeling work in this area focused on separately studying duration, magnitude or maxima of precipitation over given time scales. Since the health, economic and ecological impact of the large precipitation events depend on the combination of their characteristics, our work is a step forward towards a more realistic and practically useful modeling approach.

EO840 Room K2.31 Nash (Hybrid 07) NEW DEVELOPMENTS ON TIME SERIES MODELS Chair: Israel Martinez-Hernandez

E0409: **Statistical analysis of multi-day solar irradiance using a threshold time series model**
Presenter: **Carolina Euan**, Lancaster University, United Kingdom
The analysis of solar irradiance has important applications in predicting solar energy production from solar power plants. Although the sun provides every day more energy than we need, the variability caused by environmental conditions affects electricity production. Recently, new statistical models have been proposed to provide stochastic simulations of high-resolution data to downscale and forecast solar irradiance measurements. Most of the existing models are linear and highly depend on normality assumptions. However, solar irradiance shows strong non-linearity and is only measured during the daytime. Thus, we propose a new multi-day threshold autoregressive (TAR) model to quantify the variability of the daily irradiance time series. We establish sufficient conditions for our model to be stationary, and we develop an inferential procedure to estimate the model parameters.

E0415: **Cyclostationary processes with evolving periods and amplitudes**
Presenter: **Soumya Das**, University of Wisconsin-Madison, United States
Co-authors: Marc Genton
Wide-sense cyclostationary processes are an important class of non-stationary processes that have a periodic structure in their first- and second-order moments. The notion of cyclostationarity (in the wide sense) is extended to processes where the mean and covariance functions might depart from strict periodicities and constant amplitudes. Specifically, we propose a novel and flexible class of processes that allows periods and amplitudes of the mean and covariance functions to evolve and, therefore, accommodates a much larger class of processes than the classical cyclostationary processes. Thereafter, we investigate its properties, provide methodologies for statistical inference, and illustrate the presented methods using synthetic signals and a physical signal, from the heavens, of the magnitudes of the light emitted from the variable star R Hydrae.

E0419: **Micro-macro changepoint inference for periodic data sequences**
Presenter: **Rebecca Killick**, Lancaster University, United Kingdom
Existing changepoint approaches consider changepoints to occur linearly in time; one changepoint happens after another, and they are not linked. However, data processes may have regularly occurring changepoints, e.g. a yearly increase in ice-cream sales on the first hot weekend. Using linear changepoint approaches here will miss more global features such as a decrease in ice-cream sales in favour of sorbet. Being able to tease these global changepoint features from the more local(periodic) ones is beneficial for inference. We propose a periodic changepoint model to model this behaviour using a mixture of a periodic and linear time perspective. Built around a Reversible Jump Markov Chain Monte Carlo sampler, the Bayesian framework is used to study the local (periodic) changepoint behaviour. We integrate the local changepoint model into the pruned exact linear time (PELT) search algorithm to identify the optimal global changepoint positions. We demonstrate that the method detects both local and global changepoints with high accuracy on simulated and motivating environmental & economic applications that share periodic behaviour.

E1481: **Factor modeling of multivariate time series: A frequency components approach**
Presenter: **Raanju Sundararajan**, Southern Methodist University, United States
A frequency-domain factor model method for multivariate second-order stationary time series is proposed. The aim is to find contemporaneous linear transforms of the observed multivariate series that leads to a lower-dimensional factor series that is allowed to be multivariate stationary. Frequency components of the observed series are assumed to be linearly generated by the corresponding frequency components of a latent factor series using frequency-specific factor loading matrices. These loading matrices are then estimated using an eigendecomposition of symmetric non-negative definite matrices involving the real and imaginary parts of the spectral matrix. The factor dimension is estimated using nonparametric bootstrap tests. Consistency results concerning the estimation of eigenvalues, eigenvectors and the loading matrices are provided. The numerical performance of the proposed method is illustrated through simulation examples and an application to modeling resting-state fMRI data from autistic individuals is demonstrated.

EO302 Room K2.40 (Hybrid 08) BAYESIAN DESIGN OF EXPERIMENTS Chair: Tim Waite

E0421: **Design of experiments with functional independent variables**
Presenter: **David Woods**, University of Southampton, United Kingdom
Some novel methodology will be presented for the optimal design of experiments when at least one independent variable is a function (e.g. of time)

and can be varied continuously during a single run of the experiment. Hence, finding a design becomes a question of choosing functions to define this variation for each run in the experiment. The work is motivated by, and applied to, experiments in the pharmaceutical industry.

E0441: **A python package for Bayesian experiment design and is application in physics experiments**
Presenter: **Robert D McMichael**, National Institute of Standards and Technology, United States

The python package optbayesexpt implements optimal Bayesian experiment design for parameter estimation. A *runs good* philosophy emphasizes ease of programming, minimal demand for statistical know-how, and fast-enough execution for automation of laboratory experiments. In the code, a particle filter with typ. 10^4 particles represents the distribution of a handful of model parameters, and design setting values are chosen from a few hundred possibilities. The challenging computation of the Kullback-Liebler utility $U(d)$ is avoided using a pseudo-utility $U^*(d)$ that requires only 1D variances of forecast result distributions. Further, sampling noise is eliminated by using a single set of parameter samples to compute pseudo-utility for all candidate settings. In each measurement epoch, calculations of setting selection, likelihood of measured data, and Bayesian inference together typically require $\approx$ 10 ms of computation time. The optbayesexpt package is provided with eight example scripts and an interface module for communication with popular instrument control languages. In a laboratory demonstration, automated design with optbayesexpt reduced measurement time by a factor of 60 in magnetic resonance experiments. In simulations of Ramsey measurements (the canonical quantum measurement of energy differences), optbayesexpt outperformed published protocols.

E1384: **Gradient-based Bayesian experimental design for implicit models using mutual information lower bounds**
Presenter: **Steven Kleinegesse**, University of Edinburgh, United Kingdom
Co-authors: Michael Gutmann

A framework is introduced for Bayesian experimental design (BED) with implicit models, where the data-generating distribution is intractable but sampling from it is still possible. In order to find optimal experimental designs for such models, the approach maximizes mutual information lower bounds that are parametrized by neural networks. By training a neural network on sampled data, we simultaneously update network parameters and designs using stochastic gradient-ascent. The framework enables experimental design with a variety of prominent lower bounds and can be applied to a wide range of scientific tasks, such as parameter estimation, model discrimination and improving future predictions. Using a set of intractable toy models, we provide a comprehensive empirical comparison of prominent lower bounds applied to the aforementioned tasks. We further validate our framework on a challenging system of stochastic differential equations from epidemiology.

E1458: **Deep adaptive design: Amortizing sequential Bayesian experimental design**
Presenter: **Adam Foster**, University of Oxford, United Kingdom

The conventional approach to sequential Bayesian experimental design is to fit a posterior and optimise a design criterion at each iteration. This is computationally costly, and prevents us from using sequential design in many real-world applications such as online surveys, where we must choose each design in under a second. We will discuss Deep Adaptive Design (DAD), a new method for sequential Bayesian experimental design that does not fit posterior distributions nor optimise the criterion at each iteration of the experiment. Instead, DAD learns a design policy network that takes as input the designs and outcomes from previous iterations, and outputs the next design using a single forward pass. DAD can therefore compute the next design adaptively in milliseconds during a live experiment. The network is trained on millions of simulated experimental trajectories using a contrastive information bound as the training objective. We demonstrate that DAD learns excellent experimental design policies for a number of models, and can even outperform the conventional step-by-step approach whilst being orders of magnitude faster at deployment time.

EO886 Room K2.41 (Hybrid 09) STATISTICAL GENETICS AND THE HOST GENETICS OF COVID-19 Chair: Lloyd Elliott

E1206: **Participant powered research: Using direct to consumer genetic testing to help us understand COVID-19 host genetics.**
Presenter: **Adam Auton**, 23andMe, United States

Since its foundation in 2006, 23andMe has worked to help people access, understand, and benefit from the human genome. By enabling people to participate in scientific research, 23andMe has developed the worlds largest consented, re-contactable database for genetic research, with more than 11 million customers, a research consent rate over 80 percent, and billions of phenotypic data points. We will provide an overview of research studies conducted at 23andMe, and outline how we engage our customers in our research. We will discuss the power of the 23andMe database driving scientific discoveries that could lead to novel therapies in a wide range of diseases, and describe how this research model can be deployed for answering questions around COVID-19 host genetics.

E1265: **Genetic associations from the COVID-19 Host Genetics Initiative highlight biology behind severity and susceptibility**
Presenter: **Andrea Ganna**, Institute for Molecular Medicine, Finland, Finland

The COVID-19 Host Genetics Initiative (HGI) brings together the international human genetics community to generate, share, and analyze data to identify the genetic determinants of COVID-19 susceptibility and severity. The HGI's 6th data freeze (spring 2021) consists of 61 studies from 24 countries, including ancestries typically underrepresented in genetic studies. The meta-analysis includes 25,027 hospitalized cases and 125,548 cases with lab-confirmed or self-reported PCR-confirmed infection. We find additional genetic variation associated with severe COVID-19 symptoms. Rs35705950 at MUC5B, a strong risk variant for idiopathic pulmonary fibrosis, confers protection from severe symptoms (p 5.5e-9, OR 0.89). Novel associations to severe symptoms include a lead missense variant in surfactant protein SFTPD previously associated with COPD (p 1.9e-8, OR 1.06), and a missense variant in the lung-expressed transporter SLC22A31 which is co-expressed with surfactant protein genes (p 2.5e-8, OR 1.09). In the current analysis of infected cases, we find a strong protective effect for rs190509934 69 bases from the transcriptional start site of ACE2, a receptor for the spike protein of SARS-CoV-2 (p 3.6e-18, OR 0.69), suggesting genetic variation at ACE2 is associated with protection from SARS-CoV-2 infection. The initiative shows the power of quickly translating genetic data worldwide to biologically relevant findings

E1430: **Advances and challenges in X chromosome-aware whole genome genetic studies**
Presenter: **Lei Sun**, University of Toronto, Canada
Co-authors: Bo Chen, Radu Craiu, Wei Deng, Lloyd Elliott, Elika Garg, Andrew Paterson, Lisa Strug, Zhong Wang, Lin Zhang

The inclusion of the X chromosome (Xchr) in genome-wide association studies is known to be difficult due to multiple analytical challenges, particularly the uncertainty of X-inactivation, where one of the two Xchrs in a female may be randomly or preferentially selected to have no effect (i.e. dosage compensation), and there is also the possibility of no X-inactivation (i.e. X-inactivation escape). To date, only 0.5% of associated SNPs in the NHGRI-EBI GWAS catalog is on the Xchr, a 10-fold paucity compared to the autosomes. We will first present an Xchr association method that is robust to X-inactivation uncertainty and easy-to-implement, and compare it with other methods that have focused on X-inactivation. We will then present evidence for the previously under-appreciated phenomenon of sex differences in minor allele frequency (sdMAF), from a recent analysis of the 1000 Genomes Project data. sdMAF may affect the validity and power of the existing X-inactivation-focused association methods, as well as our current understanding of Hardy-Weinberg equilibrium and linkage disequilibrium on the Xchr. We will also discuss the relevance of these results to the Genetic Epidemiology Committee analysis of the CGEn COVID-19 Host Genome Sequence Project data. Finally, as GWAS is the basis for polygenic risk score (PRS)-based disease prediction, we will discuss opportunities and challenges facing the Xchr-inclusive PRS research.

E1695: **A Neanderthal OAS1 isoform protects individuals of European ancestry against COVID-19 susceptibility and severity**
Presenter: **Brent Richards**, McGill University, Canada

To identify circulating proteins influencing Coronavirus Disease 2019 (COVID-19) susceptibility and severity, we undertook a two-sample Mendelian randomization (MR) study, rapidly scanning hundreds of circulating proteins while reducing bias due to reverse causation and confounding. In up to 14,134 cases and 1.2 million controls, we found that an s.d. increase in OAS1 levels was associated with reduced COVID-19 death or ventilation (odds ratio (OR) = 0.54, P = 7 E- 108), hospitalization (OR = 0.61,P = 8 E-108) and susceptibility (OR = 0.78, P = 8E-106). Measuring OAS1 levels in 504 individuals, we found that higher plasma OAS1 levels in a non-infectious state were associated with reduced COVID-19 susceptibility and severity. Further analyses suggested that a Neanderthal isoform of OAS1 in individuals of European ancestry affords this protection. Thus, evidence from MR and a case-control study support a protective role for OAS1 in COVID-19 adverse outcomes. Available pharmacological agents that increase OAS1 levels could be prioritized for drug development.

EC852 Room K E. Safra (Multi-use 01) COMPUTATIONAL STATISTICS AND MACHINE LEARNING (IN-PERSON) Chair: Stanislav Nagy

E1770: **A stochastic approximation approach for parametric inference of intractable likelihood models**
Presenter: **Wentao Li**, The university of Manchester, United Kingdom

For generative models with intractable likelihood, popular parametric inference methods include the synthetic likelihood (SL), the method of simulated moments and indirect inference. These methods can be treated as simulation-based variants of the generalized method of moments. Their computational cost mainly depends on the optimization scheme and the number of pseudo dataset simulations, N, used for one estimation of the theoretical moment. In order to study the impact of N, we use the framework of the generalized empirical likelihood to study the asymptotic properties of the above methods and a simulation-based version of the empirical likelihood estimator. It is shown that optimizers given by these methods are first-order equivalent when N is fixed and as the data size goes to infinity. They are consistent and the asymptotic distribution depends on the distribution of the summary statistics. We also propose a mini-batched stochastic approximation algorithm to obtain the SL maximizer and its asymptotic variance estimator. Numerical studies show that the proposed algorithm is insensitive to the choice of N, and, compared to the commonly used synthetic-likelihood-based Metropolis-Hasting algorithm, computationally more efficient in obtaining accurate coverage probability over one order of magnitude.

E1165: **Fast computation of the angular halfspace depth**
Presenter: **Stanislav Nagy**, Charles University, Czech Republic
Co-authors: Rainer Dyckerhoff, Petra Laketa

The angular halfspace depth is a nonparametric tool for the analysis of directional data. That depth was proposed already in 1987, but its widespread use has been hampered in practice by significant computational issues. We address these problems by considering a simple projection scheme that allows reducing the computation of the angular depth to the task of evaluating a variant of the usual halfspace depth in a linear space. Efficient algorithms for exact computation and approximation of the angular halfspace depth are developed.

E1601: **Fitting higher order state-space models using hidden Markov methodology**
Presenter: **Takis Besbeas**, Athens University of Economics and Business, Greece

Count time series are frequently encountered in a variety of scientific disciplines, including ecology, biology and public health. In addition to autocorrelation, which may exceed order one, overdispersion and zero-inflation may be present in such a series. To accommodate these features, a number of researchers have proposed a flexible class of dynamic models in the state-space framework coupled with Monte Carlo Expectation Maximization (MCEM) algorithms based on the particle filter for parameter estimation. We propose a new method for model fitting based on hidden Markov model methodology. The method involves a discretisation technique of the underlying state-space together with an approach for transforming a higher-order state-space into an equivalent first-order. The proposed approach is simpler to both implement and compute, and opens the way to efficient model selection. We illustrate the practical utility of the method using an application from public health pertaining to the diagnosis coding of severe disease.

E1020: **Concentration inequalities, optimal number of layers and classification fallacy of a stochastic neural network**
Presenter: **Michele Caprio**, Duke University, United States
Co-authors: Sayan Mukherjee

Concentration inequalities are given for the output of the hidden layers of a stochastic feedforward neural network with ReLU activation, as well as for the output of the whole neural network. In addition, if the neural network is a martingale, we find a martingale inequality for the output of the hidden layers and of the whole neural network; we also identify the optimal number of layers for the neural network via an optimal stopping procedure. Finally, in the context of a two-category classification stochastic neural network, we give an approximation of the behavior of the classifier, and we give a probabilistic bound for the loss of accuracy resulting from this approximation.

EC854 Room K0.20 (Hybrid 05) METHODOLOGICAL STATISTICS AND BIOSTATISTICS Chair: Maria Brigida Ferraro

E1641: **Goodness-of-fit tests on the flat torus based on Wasserstein distance and their relevance to structural biology**
Presenter: **Javier Gonzalez-Delgado**, Toulouse Mathematics Institute and LAAS-CNRS, France
Co-authors: Alberto Gonzalez Sanz, Pierre Neuvial, Juan Cortes

The motivation comes from the study of local protein structure, which is defined by two variable dihedral angles that take values from probability distributions on the flat torus. The goal is to provide the space $\mathcal{P}(\mathbb{R}^2/\mathbb{Z}^2)$ with a metric that quantifies local structural modifications due to changes in the protein sequence, and to define associated two-sample goodness-of-fit testing approaches. Due to its adaptability to the space geometry, we focus on the Wasserstein distance as a metric between distributions. We extend existing results of the theory of Optimal Transport to the d-dimensional flat torus $\mathbb{T}^d = \mathbb{R}^d/\mathbb{Z}^d$, in particular, a Central Limit Theorem. Moreover, we assess different techniques for two-sample goodness-of-fit testing for the two-dimensional case, based on the Wasserstein distance. We provide an implementation of these approaches in R. Their performance is illustrated by numerical experiments on synthetic data and protein structure data.

E0229: **CoCoA: A conditional correlation model with association size**
Presenter: **Danni Tu**, University of Pennsylvania, United States
Co-authors: Bridget Mahony, Maxwell Bertolero, Aaron Alexander-Bloch, Theodore Satterthwaite, Danielle Bassett, Armin Raznahan, Russell Shinohara

In tasks that measure cognitive function, the trade-off between speed and accuracy requires that the two be studied together. A natural question is whether speed-accuracy coupling depends on other variables, such as sustained attention. Classical regression techniques, which make different assumptions about the covariates and outcome, are insufficient to investigate the effect of a third variable on the symmetric relationship between speed and accuracy. In response, we propose CoCoA (Conditional Correlation Model with Association Size), a statistical framework that adapts parametric and semi-parametric estimation methods inspired by genome association studies to quantify the conditional correlation as a function of additional variables. We further propose novel measures of the association size, which are analogous to effect sizes on the correlation scale, while adjusting for confounding. Using neurocognitive data from the Human Connectome Project, we demonstrate that greater sustained attention in a working memory task is associated with stronger speed-accuracy coupling while controlling for age.

E1108: **A Stein-type shrinkage limited information maximum likelihood estimator**
Presenter: **Muhammad Qasim**, Jonkoping University, Sweden

A Stein-type shrinkage estimator is proposed which is a weighted average of the ordinary least squares (OLS) and limited information maximum likelihood (LIML) estimators, with the weights inversely proportional to the Hausman test statistic. We derive the asymptotic distribution of the new estimator by means of local-to-exogenous asymptotic belief. In addition, the asymptotic risk of the Stein-type LIML estimator is calculated and shows that the risk is strictly smaller than the risk of the LIML under certain conditions. The Monte Carlo simulation and empirical application are considered to demonstrate the superiority of Stein-type LIML to the classical OLS and LIML estimators in the presence of many weak instruments and endogeneity.

CO464 Room Virtual R21 PORTFOLIO SELECTION WITH PARAMETER UNCERTAINTY Chair: Nathan Lassance

C0286: **Dynamic portfolio selection with sector-specific regularization**
Presenter: **Linqi Wang**, Universite catholique de Louvain, Belgium
Co-authors: Christian Hafner

A new algorithm is proposed for dynamic portfolio selection that takes a sector structure into account. We consider regularization with respect to within and between sector variation of portfolio weights, additional to sparsity and transaction cost controls. Our model includes two special cases as benchmarks: a dynamic conditional correlation model with shrinkage estimation of the unconditional covariance matrix, and the equally weighted portfolio. We propose an algorithm for estimation of the model parameters and calibration of the penalty terms based on cross-validation. In an empirical study, we find that the within-sector regularization contributes significantly to the reduction of out-of-sample volatility of portfolio returns. Our model improves both the pure DCC with nonlinear shrinkage and the equally-weighted portfolio out-of-sample.

C0327: **A robust approach to optimal portfolio choice with parameter uncertainty**
Presenter: **Nathan Lassance**, UCLouvain, Belgium
Co-authors: Alberto Martin-Utrera, Majeed Simaan

It is well known that estimated mean-variance portfolios deliver, on average, poor out-of-sample performance. A lesser-known fact that we characterize is that their out-of-sample performance is also very volatile. Using our analytical characterization of out-of-sample performance volatility, we propose a measure of portfolio robustness defined as the ratio between out-of-sample utility mean to out-of-sample utility standard deviation. We exploit our measure of portfolio robustness to calibrate shrinkage portfolios and show that they deliver better performance than those that ignore parameter uncertainty or only optimize out-of-sample utility mean.

C1305: **Ambiguity, learning, and equilibrium portfolio flows**
Presenter: **Alex Weissensteiner**, Free University of Bozen-Bolzano, Italy
Co-authors: Thomas Dangl, Lorenzo Garlappi

It is documented that institutional investors decrease their holdings of risky assets following both bad and good news about economic outcomes. While the reaction to bad news is consistent with documented evidence in which retail investors act as liquidity providers, the reaction of institutional investors to good news is a novel finding. We propose a general equilibrium model in which a Bayesian institution and an ambiguity averse retail investor learn about the dynamics of dividends in the economy. We show that learning about volatility and market clearing generates equilibrium patterns of portfolio flows, risk premia, and return predictability, consistent with the data

C1204: **Why naive diversification is not naive, and how to improve it**
Presenter: **Guofu Zhou**, Washington University in St. Louis, United States
Co-authors: Ming Yuan

The aim is to examine why the naive diversification is hard to beat and provides new strategies to improve it.

CO204 Room Virtual R32 UNCERTAINTY AND MODEL SELECTION IN FINANCE Chair: Mohammad Jahan-Parvar

C0243: **Model comparison with transaction costs**
Presenter: **Andrew Andrew Detzel**, University of Denver, United States
Co-authors: Mihail Velikov, Robert Novy-Marx

Failing to account for transaction costs materially impacts inferences drawn when evaluating asset pricing models, biasing tests in favour of those employing high-cost factors. Ignoring transaction costs, certain q-factor and six-factor models have high maximum squared Sharpe ratios and small alphas across 120 anomalies. They do not, however, come close to spanning the achievable mean-variance efficient frontier. Accounting for transaction costs, the Fama and French five-factor model has a significantly higher squared Sharpe ratio than either of these alternative models. At the same time, variations employing cash profitability perform better still. More generally, these results highlight the importance of incorporating real-world concerns into financial research.

C0276: **Medium- and long-term interest rate uncertainty**
Presenter: **Cisil Sarisoy**, Federal Reserve Board, United States
Co-authors: Danilo Cascaldi-Garcia

The resolution of medium- and long-term uncertainty about interest rates cause sizable expansionary economic effects that are not explained by other uncertainty sources. We construct market-based uncertainty measures for the 2- and 10-year interest rates from interest rate options. While uncertainty about the interest rate resolves around FOMCs, the dynamics of the medium- and long-term are markedly different. We show that two underlying shocks explain the joint movements between the medium- and long-term interest rate uncertainty: one which uncertainties resolve in tandem, and another where they move in opposite directions. While the first explains about 20 percent of long-term industrial production increases, the second is associated with VIX and stock market short-term movements. Other traditional uncertainty shocks, like VIX shocks, are not confounded with interest rate uncertainty shocks, highlighting the novelty of their business cycle driving forces.

C0355: **Adding fuel to the fire sales: Banks, capital regulation, and systemic risk**
Presenter: **Samuel Rosen**, Temple University, Fox School of Business, United States

In a model with heterogeneous banks and endogenous fire sales, the tightening of bank capital regulation can aggravate fire sales, leading to larger bank losses and higher systemic risk. When calibrated to the data, the least costly policies to mitigate systemic risk raise both ex-ante capital requirements and ex-post shortfall penalties. These policies also assign relatively higher capital requirements to banks that can better offset price declines during a fire sale, consistent with the recently implemented capital surcharge for global systemically important banks (G-SIBs). My findings provide further support for leading-edge macroprudential tools, including stress tests and countercyclical capital buffers.

C0488: **Foreign economic policy uncertainty and the U.S. equity returns**
Presenter: **Mohammad Jahan-Parvar**, Federal Reserve Baord of Governors, United States
Co-authors: Yuriy Kitsul, Beth Anne Wilson, Jamil Rahman

The predictive ability and economic significance of foreign economic policy uncertainty for U.S. equity returns are documented. After orthogonalizing global economic policy uncertainty (foreign EPU) with respect to the U.S. EPU, we find that it has significant predictive power for aggregate stock returns and returns of portfolios constructed on size, investment, capital expenditure, and foreign sales in 6 to 12-months ahead horizons. We additionally show that foreign EPU commands an economically significant and negative-valued premium in the cross-section of returns.

CO573 Room Virtual R33 FINANCIAL MODELLING AND FORECASTING Chair: Ekaterini Panopoulou

C0275: **The informational content of implied correlation**
Presenter: **Nikolaos Voukelatos**, University of Kent, United Kingdom
Option-implied correlation has been shown to be an efficient predictor of market returns. We examine the source of this informational content. We document that the predictive ability of implied correlation stems from the interplay between its high-frequency and low-frequency components. The high-frequency component captures short-term trends, and it is found to be a robust predictor of market returns at short horizons, outperforming the original series of implied correlations. The low-frequency component reflects longer-term trends and optimally predicts market returns at longer horizons. We also provide evidence that decomposing implied correlation substantially improves the out-of-sample predictability of market returns at horizons of up to one year.

C0828: **Structuring and pricing home equity release with alternative sharing of house price risks**
Presenter: **Jaideep Oberoi**, University of Kent, United Kingdom
Co-authors: Doug Andrews
The structure and pricing of a home-equity-release product designed for senior homeowners with a more efficient risk-sharing than traditional reverse mortgages are presented. The homeowner borrows against their home with the protection of a no-negative-equity-guarantee (NNEG), but the repayment is based on the return on a regional house-price index and a fixed premium to cover the NNEG. We illustrate the associated payoffs with the use of 20 years of home sales data from the United Kingdom (UK), alongside UK mortality data and an updated UK morbidity study. We show how the NNEG can be priced using nonparametric historical simulation, highlighting the role of basis risk in reverse mortgage pricing.

C0859: **The risk parity approach and fund of hedge funds**
Presenter: **Eirini Bersimi**, University of Kent, United Kingdom
Co-authors: Ekaterini Panopoulou, Nikolaos Voukelatos
The focus is on the risk-based approach, namely Risk Parity (RP) to asset allocation and fund of hedge funds creation. In more detail, we construct RP portfolios consisting of Hedge Funds (HF) which are competing against, the relatively new, Hedge Funds Research (HFR) RP Indices in an attempt to outperform them. The RP portfolios consist of the best 25 funds, and their performance is evaluated by typical hedge fund evaluation measures. The empirical results suggest that, for both volatility targets (10% and 12%), the constructed portfolios outperform the HFR Indices in the out-of-sample period.

C1214: **Denoising the equity premium: A wavelet quantile approach**
Presenter: **Ekaterini Panopoulou**, University of Essex, United Kingdom
Co-authors: Antonis Alexandridis
The aim is to test whether it is possible to improve point, quantile and density forecasts of equity premium returns. Previous studies have shown that a variety of economic variables fail to deliver consistently accurate out-of-sample forecasts for the equity premium. We propose a novel wavelet denoising framework in the context of equity premium forecasting. First, we decompose the time-series using wavelet analysis and then, we remove the noise in each frequency using different wavelet denoising techniques. The results show that the proposed method improves the forecasting ability of linear models indicating that wavelet denoising can successfully identify the underlying persistent signal in the equity premium. Extending our framework to wavelet quantile regression, we show our approach achieves superior point, quantile and density forecasts relative to a plethora of benchmarks. Finally, our forecasting framework survives multiple-testing control. The extensive analysis and the various robustness tests indicate that the overall out-performance of our model is not an artefact of data mining.

CO172 Room Virtual R34 INFLATION DYNAMICS Chair: Edward Knotek

C0608: **The real effects of monetary shocks: Evidence from micro pricing moments**
Presenter: **Matt Klepacz**, Federal Reserve Board, United States
Co-authors: Raphael Schoenle, Gee Hee Hong, Ernesto Pasten
The informativeness of pricing moments for monetary non-neutrality is evaluated. Empirically, the frequency of price changes is robustly informative, in line with models of price rigidities. Other moments are insignificant, or become insignificant when non-pricing moments are included. No pricing moment is a sufficient statistic. Our theoretical analysis, focused on the ratio of kurtosis over frequency of price changes in a quantitative menu cost model, finds an ambiguous relationship of the ratio with monetary non-neutrality. This result stands in contrast with existing theoretical results. We explore which assumptions explain the discrepancy, aligning theoretical and empirical results.

C0784: **Firms inflation expectations and pricing strategies during COVID-19**
Presenter: **Cristina Conflitti**, Banca d'Italia, Italy
Co-authors: Marco Bottone, Marianna Riggi, Alex Tagliabracci
The Bank of Italy's Survey on Inflation and Growth Expectations is used to explore how the COVID-19 shock affects firms pricing policies and their inflation expectations. We find that the longer the time deemed necessary to return to their normal business levels and the greater the attention they pay to their competitors pricing policies, the more likely firms are to reduce their own product prices. Moreover, firms' inflation expectations react to the expected persistence of the macroeconomic effects of the shock. We rationalize this evidence through the lens of a general equilibrium model.

C1261: **What can stockouts tell about inflation: Evidence from online micro data**
Presenter: **Oleksiy Kryvtsov**, Bank of Canada, Canada
Co-authors: Alberto Cavallo
A detailed micro dataset on product availability is used to construct a direct high-frequency measure of consumer product shortages during 2020–2021 pandemic. We document a widespread multi-fold rise in shortages in nearly all sectors early in the pandemic. Over time, the composition of shortages evolved from many temporary stockouts to mostly discontinued products, concentrated in fewer sectors. We show that product shortages have significant but transitory inflationary effects, and that these effects can be associated with the elevated cost of replenishing inventories.

C0535: **The effects of price endings on price rigidity: Evidence from VAT changes**
Presenter: **Edward Knotek**, Federal Reserve Bank of Cleveland, United States
Micro price data underlying the CPI in Israel reveal that most stores have a favored price ending–a final digit to the right of the decimal, usually zero or nine, used for a large majority of prices. Using VAT rate changes as exogenous shocks that affect prices regardless of their ending digit, we find that the frequency of adjustment for non-favoured price endings increases by twice as much as the frequency of adjustment for favored endings, consistent with favored endings playing a causal role in generating price rigidity. In the aggregate, favored endings produce sluggish pass-through of VAT rate changes.

CO468 Room Virtual R38 MACHINE LEARNING TECHNIQUES, CLIMATE CHANGE AND PORTFOLIO SELECTION Chair: Jose Olmo

C0623: **Long-term climate forecasts: A future heterogenous warming**
Presenter: **Jesus Gonzalo**, Universidad Carlos III de Madrid, Spain

Co-authors: Lola Gadea

Climate is a long-term issue and therefore climate forecasts should be long-run forecasts. They are crucial in order to design the mitigation policies required to fulfil one of the main objectives of the Paris Climate Agreement (PCA): to limit the long-term mean temperature increase to well below 1.5-2C above the pre-industrial period levels (1850-1900). Using a realized quantile methodology where quantiles are converted into time series objects, a simple method is proposed to produce long-term temperature density forecasts from observational data. Analyzing the observational data from global cross-section stations CRU 1880-2018 we obtain three sets of important results: (i) In 10-25 years the mean global temperature will be above the 2C degrees upper bound set by the PCA and by the end of the 21st century the increment will be of 3.5C-4C degrees above pre-industrial levels; (ii) this increase is larger in the lower quantiles (e.g. q05 will go from the 0.07C pre-industrial level to 2.07C in 25 years and to 4.06C in 100 years) than in the upper quantiles (e.g. q95 from 25.6C to 27.05C in 25 years and to 27.66C in 100 years) producing a decrease in the variance of the temperature distribution (more serious consequences than the standard increase in the mean), and (iii) there is a clear accelerating warming heterogeneous process.

C0598: **A network regression model with an estimated interaction matrix**

Presenter: **Jose Olmo**, University of Southampton, United Kingdom

Co-authors: Marcos Sanso-Navarro

A network regression model is proposed that incorporates exogenous neighboring effects into standard cross-sectional specifications. The interaction matrix is given by realizations of a functional coefficient that captures the network effects between the neighboring covariates and the outcome variable. This matrix is estimated using sieve regression methods and a Taylor expansion about a grid of reference points spanning the support of the distance variable that establishes the similarity between observations. The standardized estimator of the functional coefficient follows a zero-mean Gaussian process and the associated network parameter estimates are consistent and asymptotically normal. We also implement a uniform test to statistically assess for the presence of network effects. The empirical application studies environmental Engel curves, recently discussed in the literature, and finds strong evidence of neighboring effects in the relationship between households' income and the amount of pollution embodied in the goods and services they consume.

C0602: **Machine learning the carbon footprint of Bitcoin mining**

Presenter: **Hector Calvo-Pardo**, University of Southampton, United Kingdom

Co-authors: Tullio Mancini, Jose Olmo

Building on an economic model of rational Bitcoin mining, we measure the carbon footprint of Bitcoin mining power consumption using feed-forward neural networks. We find associated carbon footprints of 2.77, 16.08, and 14.99 MtCO2e for 2017, 2018, and 2019 based on a novel bottom-up approach, which (i) conform with recent estimates, (ii) lie within the economic model bounds while (iii) delivering much narrower prediction intervals, and yet (iv) raise alarming concerns, given recent evidence (e.g., from climate-weather integrated models). By 2024, conservative point forecasts based on an exponential trend found for the network hash rate suggest a carbon footprint of 132.01 MtCO2e, similar to the combined annualized 2019 greenhouse gas emissions of Belgium (100 MtCO2e) and Denmark (32 MtCO2e). We demonstrate how machine learning methods can contribute to non-for-profit pressing societal issues, like global warming, where data complexity and availability can be overcome.

C0678: **Portfolio selection under systemic risk: A QRNN-based approach**

Presenter: **Abderrahim Taamouti**, Durham University Business School, United Kingdom

Co-authors: Weidong Lin

The aim is to improve the traditional mean-variance (MV) portfolio selection model by accounting for systemic risk and using machine learning techniques. Our objective is to formulate the portfolio selection as a three-step supervised learning problem, which allows for considering systemic risk when constructing optimal portfolios. In the first step, we use a quantile regression neural network (QRNN) to predict conditional quantiles for stock returns. Based on the obtained quantiles, we estimate the marginal distributions for individual assets and the market portfolio. In the second step, we use copula to model the dependence structure among assets and generate return scenarios. Lastly, we solve the portfolio optimization problem dynamically by maximizing a conditional Sharpe ratio (CoSR) based on the simulated return scenarios. Thereafter, we run several comparative studies using real data on big US financial institutions. The backtesting results demonstrate the superiority of our proposed portfolio over other benchmark portfolios. In particular, we compare the out-of-sample performance of our portfolio with those of: (i) a portfolio that maximizes the unconditional Sharpe ratio (SR); (ii) a Global Minimum Variance Portfolio (GMVP); and (iii) an equally weighted portfolio $(1/N)$.

CO180 Room Virtual R39 ROBUSTNESS IN TIME SERIES Chair: Pascal Bondon

C1260: **Robust testing for correlation in a non-i.i.d. setting**

Presenter: **Yufei Li**, Queen Mary University of London, United Kingdom

Co-authors: Liudas Giraitis

The methodology of testing for correlation and cross-correlation is extended to a wider class of data and the finite sample performance of the robust testing procedures are studied. Models with non-smooth deterministic and stochastic (unit root type) scale factors, which are not covered in previous research, are studied theoretically and in the Monte Carlo experiments to compute and compare the size of both standard and robust tests. In the Monte Carlo study, we use models that include deterministic and stochastic scale factors mentioned above, demonstrating the performance of the robust statistics for testing for cross-correlation. In an empirical exercise, we test for autocorrelation and cross-correlation for some financial data to show the applications of the robust testing method. By comparing the results based on the standard statistics and the robust ones, the advantages of the robust testing procedures in empirical research are further uncovered and emphasized.

C1268: **Robust estimation of volatility models in the presence of additive outliers**

Presenter: **Luiz Hotta**, University of Campinas, Brazil

Co-authors: Jean Sabino Diniz, Eduardo Gabriel Pinheiro, Carlos Trucios

Estimation and prediction of volatility in univariate and multivariate financial time series are of crucial importance. One of the features of data from finance is the variety of types of series and applications, leading to several models proposed in the literature. Another important feature is the presence of outliers, especially additive outliers. We discuss the estimation of the volatility using several models, from the simple univariate GARCH model to the high dimensional cDCC model. For each of the entertained models, we first present the effects of the outliers on the estimates from traditional non-robust methods. Then, we propose a robust estimator and compare the performance of the traditional and robust methods. The comparison considers different frequencies and sizes of the outliers.

C1338: **Blind source separation based on M autocovariance matrices**

Presenter: **Sara Taskinen**, University of Jyvaskyla, Finland

Co-authors: Klaus Nordhausen, David Tyler

Assume that the observed p time series are linear combinations of p latent uncorrelated weakly stationary time series. The aim of blind source separation (BSS) is to find an estimate for the unmixing matrix which transforms the observed time series back to uncorrelated latent time series. Classical AMUSE (Algorithm for Multiple Unknown Signals Extraction) method solves the BSS problem by jointly diagonalizing the sample covariance matrix and the sample autocovariance matrix with chosen lag. A natural extension of AMUSE is SOBI (Second Order Blind Identifica-

tion) method, which approximately jointly diagonalizes the sample covariance matrix and several sample autocovariance matrices with chosen lags to solve the unmixing matrix. It is well known that in the presence of outliers, the sample covariance matrix and sample autocovariance matrices perform poorly and yield unreliable unmixing matrix estimates. We propose a robust blind source separation method that utilizes so-called M autocovariance matrices. The M autocovariance matrices are similar to the classical M estimators in that they downweight the outliers using some preselected, bounded weight function. Simulation studies and a real data example are used to illustrate robustness and efficiency properties of proposed methods.

C1320: **A robust longitudinal study of the influence of air pollutants on children**

Presenter: **Pascal Bondon**, CentraleSupelec, France

Co-authors: Ian Danilevicz, Valderio Anselmo Reisen, Faradiba Sarquis

To measure the influence of environmental predictors on allergic children, we combine a robust principal component analysis (RPCA) and a robust estimation of a linear mixed model (LMM) in a longitudinal study. This strategy is justified by the verification that robust estimation of a LMM might not deal with correlated outliers in covariates. RPCA transforms the original covariates into uncorrelated variables which are used as new covariates in the LMM. In the real data analysis, RPCA exhibits three principal components mainly related with humidity and particulates matter with a diameter smaller than 10 micrometers (PM10) and 2.5 micrometers (PM2.5), respectively. The study shows that high levels of total immunoglobulin E, dry weather and PM10 are significant risk factors for children's respiratory diseases.

CO694 Room Virtual R40 ASSET PRICING II Chair: Benjamin Holcblat

C1625: **What do the portfolios of individual investors reveal about the cross-section of equity returns**

Presenter: **Laurent Calvet**, EDHEC, France

Co-authors: Sebastien Betermier, Samuli Knupfer, Jens Kverner

A parsimonious set of equity factors is constructed by sorting stocks according to the sociodemographic characteristics of the individual investors who own them. The analysis uses administrative data on the stockholdings of Norwegian investors in 1997-2018. Consistent with financial theory, a mature-minus-young factor, a high wealth-minus low wealth factor, and the market factor price stock returns. Our three factors span size, value, investment, profitability, and momentum, and perform well in out-of-sample bootstrap tests. The tilts of investor portfolios toward the new factors are driven by wealth, indebtedness, macroeconomic exposure, age, gender, education, and investment experience. Our results are consistent with hedging and sentiment jointly driving portfolio decisions and equity premia.

C1511: **Measuring corporate bond market dislocations**

Presenter: **Nina Boyarchenko**, Federal Reserve Bank of New York, United States

The Corporate Bond Market Distress Index (CMDI) is proposed to quantify corporate bond market dislocations in real-time. The index takes a preponderance-of-metrics perspective to combine a broad set of measures of market functioning from primary and secondary markets but not driven by any one statistic. We document that the index correctly identifies periods of dislocations and predicts future realizations of commonly used measures of market distress, while the converse is not the case. Moreover, the CMDI is an economically and statistically significant predictor of future economic activity, even after controlling for standard predictors, including credit spreads.

C1510: **Test assets and weak factors**

Presenter: **Stefano Giglio**, Yale and NBER, United States

Estimation and testing of factor models in asset pricing require choosing a set of test assets. The choice of test assets determines how well different factor risk premia can be identified: if only a few assets are exposed to a factor, that factor is weak, which makes standard estimation and inference incorrect. In other words, the strength of a factor is not an inherent property of the factor: it is a property of the cross-section used in the analysis. We propose a novel way to select assets from a universe of test assets and estimate the risk premium of a factor of interest, as well as the entire stochastic discount factor, that explicitly accounts for weak factors and test assets with highly correlated risk exposures. We refer to our methodology as supervised principal component analysis (SPCA), because it iterates an asset selection step and a principal-component estimation step. We provide the asymptotic properties of our estimator, and compare its limiting behavior with that of alternative estimators proposed in the recent literature, which rely on PCA, Ridge, Lasso, and Partial Least Squares (PLS). We find that the SPCA is superior in the presence of weak factors, both in theory and in finite samples. We illustrate the use of SPCA by applying it to estimate the risk premia of several tradable and nontradable factors, to evaluate asset manager's performance, and to de-noise asset pricing factors.

C0166: **Discussant**

Presenter: **Benjamin Holcblat**, University of Luxembourg, Luxembourg

Based on the session speakers' recent work, relevant progress in the field of asset pricing will be discussed.

Authors Index

