PROGRAMME AND ABSTRACTS

14th International Conference on

Computational and Financial Econometrics (Virtual CFE 2020)

http://www.cfenetwork.org/CFE2020

and

13th International Conference of the
ERCIM (European Research Consortium for Informatics and Mathematics) Working Group on

Computational and Methodological Statistics (Virtual CMStatistics 2020)

http://www.cmstatistics.org/CMStatistics2020

19 – 21 December 2020

ISBN 978-9963-2227-9-7

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

CFE 2020 Co-chairs:
Anurag Banerjee, Scott Brave, Peter Pedroni and Mike So.

CFE 2020 Programme Committee:

Knut Are Aastveit, Alessandra Amendola, David Ardia, Josu Arteche, Anindya Banerjee, Travis Berge, Monica Billio, Raffaella Calabrese, Massimiliano Caporin, Julien Chevallier, Serge Darolles, Luca De Angelis, Filippo Ferroni, Ana-Maria Fuertes, Massimo Guidolin, Harry Haupt, Masayuki Hirukawa, Benjamin Holcblat, Rustam Ibragimov, Laura Jackson Young, Michel Juillard, Edward Knotek, Robinson Kruse-Becher, Svetlana Makarova, Ilia Negri, Ingmar Nolte, Jose Olmo, Yasuhiro Omori, Jesus Otero, Michael Owyang, Alessia Paccagnini, Indeewara Perera, Jean-Yves Pitarakis, Tommaso Proietti, Artem Prokhorov, Tatevik Sekhposyan, Etsuro Shioji, Michael Smith, Robert Taylor, Martin Wagner and Ralf Wilke.

CMStatistics 2020 Co-chairs:
Tapabrata Maiti, Sofia Olhede, Michael Pitt, Cheng Yong Tang and Tim Verdonck.

CMStatistics 2020 Programme Committee:

Julyan Arbel, Raffaele Argiento, Andreas Artemiou, Arnab Bhattacharjee, Yuguo Chen, Radu Craiu, Marzia A Cremona, Mike Daniels, Peng Ding, Yuexiao Dong, Fabrizio Durante, Marco Ferreira, Konstantinos Fokianos, Irina Gaynanova, Jeff Goldsmith, Gil Gonzalez-Rodriguez, Alessandra Guglielmi, Jan Hannig, Christopher Hans, Jiashun Jin, Kshitij Khare, Claudia Kirch, Olga Klopp, Thomas Kneib, Keith Knight, Eric Kolaczyk, Yoonkyung Lee, Chenlei Leng, Christophe Ley, Zeda Li, Zhigang Li, George Michailidis, Ursula Mueller, Kalliopi Mylona, Hernando Ombao, Anna K Panorska, Agnese Panzera, Marianna Pensky, Elisa Perrone, Mario Peruggia, Denys Pommeret, Igor Pruenster, Peter Rousseeuw, Juan Eloy Ruiz-Castro, Rajen Shah, Sanjoy Sinha, Dongchu Sun, Masayuki Uchida, Mattias Villani, HaiYing Wang, Tongtong Wu, Tingting Zhang, Wenyang Zhang, Xiaoke Zhang, Jiwei Zhao, Ricardas Zitikis and Xavier de Luna.

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 virtually in these difficult times. This year we are passing through extraordinary events that are significantly affecting our personal and professional lives. Until September 2020 we were still planning to hold part of the conference in-person. However due to the Covid-19 pandemic we were forced to have the whole conference hosted virtually. Despite the organization challenges caused by these changes in such a short time, we are happy to welcome warmly the over 1100 participants. More than ever, we acknowledge the efforts of all those involved in the conference, especially the session organizers, who had to rebuild their sessions to adapt to the online requirements.

The 14th International Conference on *Computational and Financial Econometrics* (CFE 2020) and the 13th International Conference of the ERCIM Working Group on *Computational and Methodological Statistics* (CMStatistics 2020) 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 2020 programme consists of 250 sessions, five plenary talks and about 1000 presentations. There are over 1100 participants.

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 virtual 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 Elsevier journal Econometrics and Statistics (EcoSta) has been 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 2021 will be held at King's College London, from Saturday the 18th of December 2021 to Monday the 20th of December 2021. Tutorials will take place on Thursday the 17th of December 2021. You are invited and encouraged to participate in these events actively.

We wish you a productive and stimulating virtual 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 1900 members and the following specialized teams

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

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

CFEnetwork COMPUTATIONAL AND FINANCIAL ECONOMETRICS

http://www.CFEnetwork.org

The Computational and Financial Econometrics (CFEnetwork) comprises a number of specialized teams in various research areas of theoretical and applied econometrics, financial econometrics and computation, and empirical finance. The teams contribute to the activities of the network 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)

2020-12-19

A - Keynote
CFE - CMStatistics
08:45 - 09:45

B
CFE - CMStatistics
09:55 - 11:35

C - Keynote
CFE - CMStatistics
11:45 - 12:35

Networking Lunch Break
12:35 - 13:35

D
CFE - CMStatistics
13:35 - 15:15

E
CFE - CMStatistics
15:25 - 17:30

F
CFE - CMStatistics
17:40 - 18:55

Virtual Welcome Reception
19:00 - 20:00

2020-12-20

G
CFE - CMStatistics
08:45 - 10:50

H
CFE - CMStatistics
11:00 - 12:15

Networking Lunch Break
12:15 - 13:35

I
CFE - CMStatistics
13:15 - 14:55

J
CFE - CMStatistics
15:05 - 16:20

K
CFE - CMStatistics
16:30 - 18:10

L - Keynote
CFE - CMStatistics
18:20 - 19:10

2020-12-21

M
CFE - CMStatistics
08:45 - 10:00

N
CFE - CMStatistics
10:10 - 12:15

Networking Lunch Break
12:15 - 13:15

O - Keynote
CFE - CMStatistics
13:15 - 14:05

P
CFE - CMStatistics
14:15 - 15:55

Q
CFE - CMStatistics
16:05 - 17:45

R - Keynote
CFE - CMStatistics
17:55 - 18:55

VIRTUAL TUTORIALS, MEETINGS AND SOCIAL EVENTS

TUTORIALS

Tutorials will take place on Friday the 18th of December 2020. The first tutorial ("Forecasting after breaks") will be delivered by Prof. Sir David Hendry, and Dr Jennifer L. Castle, University of Oxford, UK, 9:00-13:30 (GMT). The second tutorial ("Bayesian modeling of brain imaging data") will be delivered by Prof. Michele Guindani, University of California, Irvine, USA, 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* meetings will take place on Friday the 18th of December 2020, 16:00-16:50 (GMT).
- The *CSDA and Annals of Statistical Data Science Editorial Board* will take place on Friday the 18th of December 2020, 17:00-17:50 (GMT).

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

ACCESS TO THE VIRTUAL CONFERENCE

- Your access to the virtual conference is personal and cannot be transferred to anybody else. If you share your credentials and any non-registered participant enters the meeting with them, they will be banned.
- Keep at hand your registration number. You can find your registration number in the email confirming that you are registered or in the conference receipt that you can download from the registration tool. The conference staff may request this number for identification purposes.

Scientific programme and social events

- The conference is live streaming, and it will not be recorded. The oral presentations will take place through Zoom, while the social events and poster presentations will run in Gather Town.
- **Scientific programme:** The virtual sessions are accessible from the interactive schedule. The conference programme time is set in GMT. Indications to access the rooms can be found on the website.
- **Networking lunch breaks:** During the 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 virtual welcome reception for registered participants will take place on Saturday 19th of December 2020, 19:00-20:00 (GMT). Indications to access the networking space can be found on the website.

Presentation instructions

The paper presentations will take place through Zoom. Speakers should install the application, have a stable internet connection, and make sure their video and audio is working. They will share their slides when the chair requires it, present their talk, and be ready to answer the question after the presentation. Detailed indications for speakers can be found on the website. 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 17th 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 by the name Angel followed by the number of the room, will assist in giving the rights to participate as the chair requests it. If any speaker is missing or has a technical problem, the chair can pass to the next speaker and come back later to resume if possible. Detailed indications for the session chairs can be found on the website.

Test sessions

Two test sessions will be set up for Sunday the 13th of December 2020: from 09:00 to 10:00 and from 16:00 to 17:00 GMT. The participants will be able to enter any of the virtual rooms in the programme to test their presentations, video, micro and audio. Detailed indications for the test sessions can be found on the website.

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

General Information **I**

Committees . . . III
Welcome . . . IV
CMStatistics: ERCIM Working Group on Computational and Methodological Statistics . . . V
CFEnetwork: Computational and Financial Econometrics . . . V
Scientific programme . . . VI
Tutorials, Meetings and Social events . . . VII
Access to the virtual content . . . VII
Publications outlets of the journals EcoSta and CSDA and Call for papers . . . VIII

Keynote Talks **1**

Opening and keynote talk 1 (David Hendry, University of Oxford, United Kingdom) Saturday 19.12.2020 at 08:45 - 09:45
Econometric methods for empirically modelling climate change . . . 1
Keynote talk 2 (Enno Mammen, Heidelberg University, Germany) Saturday 19.12.2020 at 11:45 - 12:35
Additive models: Smooth backfitting, high dimensions, random mixed forests . . . 1
Keynote talk 3 (Valentina Corradi, University of Surrey, United Kingdom) Sunday 20.12.2020 at 18:20 - 19:10
Conditional quantile coverage: An application to growth at risk . . . 1
Keynote talk 4 (Timo Terasvirta, Aarhus University, Denmark) Monday 21.12.2020 at 13:15 - 14:05
Four Australian banks and the multivariate time-varying smooth transition correlation GARCH model . . . 1
Keynote talk 5 and closing (Michele Guindani, University of California, Irvine, United States) Monday 21.12.2020 at 17:55 - 18:55
Within, between, beyond: Methods for assessing variability in brain imaging . . . 1

Parallel Sessions **2**

Parallel Session B – CFE-CMStatistics (Saturday 19.12.2020 at 09:55 - 11:35) **2**

EO245: Recent advances in functional data analysis (Room: R11) . . . 2
EO159: Advances in directional statistics (Room: R12) . . . 2
EO662: Estimation methods for extreme events (Room: R13) . . . 3
EO638: Non-parametric analysis of complex data (Room: R15) . . . 3
EO223: Projection pursuit (Room: R16) . . . 4
EO077: Advances in copula theory (Room: R18) . . . 4
EO199: Computational and theoretical statistics for stochastic processes (Room: R19) . . . 5
EO538: Analyzing compositional, distributional and relative abundance data I (Room: R20) . . . 5
EO480: Bayesian computation (Room: R22) . . . 6
EO183: The Stein method and statistics (Room: R24) . . . 7
EO594: Asymptotic theory in statistics (Room: R25) . . . 7
EC792: Contributions in spatial statistics (Room: R14) . . . 8
EC790: Contributions in survival analysis (Room: R21) . . . 8
CO037: Topics in time series econometrics (Room: R02) . . . 9
CO065: High frequency data in economics and finance (Room: R03) . . . 9
CO063: Advances in robust estimation and inference: Theory and applications I (Room: R04) . . . 10
CO133: Regularization and network approaches in financial applications (Room: R06) . . . 11
CO295: Advances in credit risk modelling (Room: R07) . . . 11
CO219: Statistical learning of high dimensional data (Room: R17) . . . 12
CC809: Contributions in monetary policy (Room: R08) . . . 12

Parallel Session D – CFE-CMStatistics (Saturday 19.12.2020 at 13:35 - 15:15) **14**

EO057: Statistics for Hilbert spaces I (Room: R11) . . . 14
EO738: Recent advances in functional time series (Room: R12) . . . 14
EO508: Advances in financial and pandemic econometrics (Room: R13) . . . 15
EO131: New proposals for the analysis of ordinal and mixed-type data (Room: R14) . . . 15
EO285: Data science methods for intelligent text and video processing (Room: R15) . . . 16
EO532: Network diffusion modelling and stochastic optimization (Room: R16) . . . 17
EO041: Advances in high dimensional time series models (Room: R17) . . . 17
EO542: New advances in biomedical data analysis (Room: R18) . . . 18
EO766: Recent advances in statistical methods for mobile health (Room: R19) . . . 18
EO552: Analyzing compositional, distributional and relative abundance data II (Room: R20) . . . 19
EO105: Recent advance is network analysis (Room: R21) . . . 20
EO684: Bayesian inverse problems (Room: R22) . . . 21
EO055: Track: Bayesian semi- and non-parametric methods II (Room: R23) . . . 21
EO059: Advances in Monte Carlo computation for data sciences (Room: R24) . . . 22
EO782: Insurance analytics (Room: R25) . . . 22
CO231: Topics in statistical learning and time series econometrics (Room: R02) . . . 23
CO307: Advances in financial econometrics (Room: R03) . . . 23
CO255: Modelling, forecasting, volatility and accuracy (Room: R04) . . . 24
CO305: Quantitative management (Room: R06) . . . 25

CO067: CLIMATE FINANCE (Room: R07) . . . 25
CO303: UNCERTAINTY IN EMPIRICAL MACROECONOMICS (Room: R08) . . . 26

Parallel Session E – CFE-CMStatistics (Saturday 19.12.2020 at 15:25 - 17:30) **27**
EO600: FUNCTIONAL DATA AND COMPLEX DATA ANALYSIS (Room: R11) . . . 27
EO277: RECENT DEVELOPMENTS ON ROBUST FUNCTIONAL DATA ANALYSIS (Room: R12) . . . 27
EO139: RECENT ADVANCES IN STATISTICS OF EXTREMES (Room: R13) . . . 28
EO632: RECENT ADVANCES TOWARD UNDERSTANDING DEEP LEARNING (Room: R14) . . . 29
EO173: TRENDS IN THE ANALYSIS OF LARGE AND COMPLEX DATA (Room: R15) . . . 30
EO247: RECENT ADVANCES IN NON-GAUSSIAN STOCHASTIC PROCESSES (Room: R16) . . . 30
EO484: TOPICS IN HIGH-DIMENSIONAL STATISTICAL INFERENCE (Room: R17) . . . 31
EO492: COPULAS AND DEPENDENCE MODELLING (Room: R18) . . . 32
EO369: RECENT DEVELOPMENT IN NETWORK ANALYSIS AND CLUSTER ANALYSIS (Room: R19) . . . 32
EO075: RECENT ADVANCES IN STOCHASTIC NETWORK MODELS (Room: R20) . . . 33
EO740: BIOSTATISTICS IN CANCER RESEARCH (Room: R21) . . . 33
EO514: ADAPTIVE BAYESIAN METHODS FOR TIME AND SPATIAL SERIES ANALYSIS (Room: R22) . . . 34
EO279: ADVENTURES IN BAYESIAN NONPARAMETRICS (Room: R23) . . . 35
EO546: CHALLENGES AND RECENT ADVANCES IN OPTIMAL DESIGN (Room: R24) . . . 36
EO744: RECENT DEVELOPMENT OF SUFFICIENT MULTIVARIATE METHODS (Room: R25) . . . 36
EC800: CONTRIBUTIONS IN METHODOLOGICAL STATISTICS (Room: R02) . . . 37
CI025: ECONOMETRIC CHALLENGES CAUSED BY PANDEMIC (Room: R04) . . . 38
CO113: REGIME CHANGE I: BUSINESS CYCLES AND REGIME CHANGE (Room: R03) . . . 38
CO720: TOPICS OF MACHINE LEARNING AND ECONOMETRICS IN MONETARY POLICY (Room: R07) . . . 39
CO299: LOCAL PROJECTIONS AND APPLICATIONS (Room: R08) . . . 40
CC816: CONTRIBUTIONS IN ECONOMETRIC MODELLING (Room: R06) . . . 40

Parallel Session F – CFE-CMStatistics (Saturday 19.12.2020 at 17:40 - 18:55) **42**
EO690: FUNCTIONAL DATA DEFINED OVER ARBITRARILY SHAPED DOMAINS (Room: R11) . . . 42
EO211: ADVANCES IN SPORTS (Room: R13) . . . 42
EO512: NON-STANDARD STATISTICS ON COMPLEX DATA (Room: R14) . . . 42
EO528: TOPICS IN CAUSAL INFERENCE: SELECTION BIAS AND SENSITIVITY ANALYSIS (Room: R15) . . . 43
EO502: RECENT ADVANCES IN CAUSAL INFERENCE (Room: R16) . . . 43
EO526: HIGH-DIMENSIONAL INFERENCE FOR COMPLEX PROBLEMS (Room: R17) . . . 44
EO494: STATISTICAL AND MACHINE LEARNING METHODOLOGY FOR MEDICAL IMAGING (Room: R18) . . . 44
EO155: STATISTICAL METHODS FOR IMAGING DATA ANALYSIS (Room: R19) . . . 45
EO151: STATISTICAL ANALYSIS OF MULTIPLE NETWORKS (Room: R20) . . . 45
EO177: SURVIVAL ANALYSIS (Room: R21) . . . 46
EO141: STATISTICAL MODELING OF COVID-19 PANDEMIC (Room: R22) . . . 46
EO241: RECENT ADVANCES IN BAYESIAN METHODS FOR CORRELATED DATA (Room: R23) . . . 47
EO580: PLANNED AND UNPLANNED PRESENCE IN OBSERVATIONAL RESEARCH (Room: R24) . . . 47
EO257: RECENT ADVANCES IN ESTIMATION THEORY (Room: R25) . . . 48
CO273: TIME SERIES AND FORECASTING (Room: R02) . . . 48
CO095: SENTIMENTS, UNCERTAINTY AND MACHINE LEARNING (Room: R04) . . . 49
CC808: CONTRIBUTIONS IN MACROECONOMETRICS (Room: R08) . . . 49
CG028: CONTRIBUTIONS IN APPLIED ECONOMETRICS I (Room: R06) . . . 50

Parallel Session G – CFE-CMStatistics (Sunday 20.12.2020 at 08:45 - 10:50) **51**
EO373: SOME RECENT ADVANCES IN MIXTURE AND CLUSTER ANALYSES (Room: R12) . . . 51
EO205: STATISTICS FOR COMPLEX RANDOM SYSTEMS: THEORY AND PRACTICE (Room: R13) . . . 51
EO548: RECENT ADVANCES IN SPATIAL AND TIME SERIES MODELS (Room: R14) . . . 52
EO121: PIONEERING NEW FRONTIERS IN DISTRIBUTION AND MODELLING (Room: R16) . . . 53
EO325: MODEL VALIDATION (Room: R18) . . . 53
EO087: ADVANCES IN SURVIVAL AND RELIABILITY I (Room: R21) . . . 54
EC786: CONTRIBUTIONS IN COMPUTATIONAL STATISTICS (Room: R08) . . . 54
EC449: CONTRIBUTIONS IN MULTIVARIATE STATISTICS (Room: R11) . . . 55
EC796: CONTRIBUTIONS IN BAYESIAN STATISTICS (Room: R22) . . . 56
EG084: CONTRIBUTIONS IN CAUSAL INFERENCE AND GRAPHICAL MODELS (Room: R15) . . . 57
EG074: CONTRIBUTIONS IN TIME SERIES (Room: R17) . . . 57
EG010: CONTRIBUTIONS IN APPLIED STATISTICS I (Room: R20) . . . 58
CI021: ADVANCES IN BAYESIAN ANALYSIS AND APPLICATIONS (Room: R04) . . . 59
CO454: MODELLING COMPLEX TIME SERIES IN ECONOMICS AND FINANCE (Room: R02) . . . 60
CO107: REGIME CHANGE II: FINANCE, MACRO, POLICY REGIMES (Room: R07) . . . 60
CC802: CONTRIBUTIONS IN FINANCIAL ECONOMETRICS I (Room: R03) . . . 61
CG066: CONTRIBUTIONS IN HIGH FREQUENCY DATA IN ECONOMICS AND FINANCE (Room: R06) . . . 62

Parallel Session H – CFE-CMStatistics (Sunday 20.12.2020 at 11:00 - 12:15) **63**
EO706: FRONTIERS IN POPULATION GENETICS (Room: R08) . . . 63
EO421: EXPERIMENTAL DESIGN AND DATA ANALYSIS (Room: R11) . . . 63
EO169: CLUSTERING OF COMPLEX DATA STRUCTURE (Room: R12) . . . 63
EO498: PUBLIC POLICY ANALYSIS AND MACHINE LEARNING I (Room: R13) . . . 64
EO650: DATA-CENTRIC ENGINEERING: A NEW CHALLENGE FOR STATISTICIANS (Room: R14) . . . 64
EO668: MODELING UNCERTAINTY AND VAGUENESS IN DECISION MAKING AND ECONOMICS (Room: R15) . . . 65
EO081: ADVANCES IN SURVIVAL AND RELIABILITY II (Room: R21) . . . 65
EO115: STATISTICS AND DECISION MAKING FOR THE COVID-19 PANDEMIC (Room: R22) . . . 66
EC799: CONTRIBUTIONS IN APPLIED STATISTICS II (Room: R20) . . . 66
EG012: CONTRIBUTIONS IN COMPUTATIONAL AND METHODOLOGICAL STATISTICS (Room: R16) . . . 67
CO031: TOPICS IN TIME SERIES AND PANEL DATA ECONOMETRICS (Room: R02) . . . 67
CO191: FINANCIAL ECONOMETRICS: INTRINSIC TIME, VOLATILITY ESTIMATION, JUMP TESTING (Room: R03) . . . 68
CO694: CRYPTOCURRENCY ANALYTICS (Room: R04) . . . 68
CC814: CONTRIBUTIONS IN APPLIED ECONOMETRICS II (Room: R06) . . . 69
CG747: CONTRIBUTIONS IN EMPIRICAL MACROECONOMICS (Room: R07) . . . 69

Parallel Session I – CFE-CMStatistics (Sunday 20.12.2020 at 13:15 - 14:55) **71**
EO696: INFERENCE FOR FUNCTIONAL PARAMETERS (Room: R11) . . . 71
EO604: STATISTICAL PROBLEMS UNDER PRIVACY CONSTRAINTS (Room: R12) . . . 71
EO283: RECENT ADVANCES IN EXTREME VALUE ANALYSIS (Room: R13) . . . 72
EO478: RECENT ADVANCES FOR THE MODELLING OF COMPLEX DATA (Room: R14) . . . 72
EO482: RECENT ADVANCES IN MULTIVARIATE AND MULTI-VARIABLE ANALYSIS (Room: R15) . . . 73
EO381: ADVANCES IN CAUSAL INFERENCE (Room: R16) . . . 73
EO468: RECENT DEVELOPMENT IN STATISTICAL ANALYSIS OF NETWORK DATA (Room: R17) . . . 74
EO530: METHODOLOGICAL AND COMPUTATIONAL ADVANCES IN COPULA MODELS (Room: R18) . . . 75
EO752: ANALYZING AND PREDICTING DIFFERENT OUTCOMES IN CLINICAL STUDIES (Room: R20) . . . 75
EO652: RECENT ADVANCES IN SEMIPARAMETRIC SURVIVAL ANALYSIS (Room: R21) . . . 76
EO486: FLEXIBLE BAYESIAN MODELS FOR COMPLEX DATA (Room: R23) . . . 76
EO688: ADVANCES IN MULTIVARIATE BAYESIAN METHODS (Room: R24) . . . 77
EC798: CONTRIBUTIONS IN STOCHASTIC PROCESSES (Room: R22) . . . 78
EC801: CONTRIBUTIONS IN NONPARAMETRIC STATISTICS AND RESAMPLING (Room: R25) . . . 78
CO427: ASSET PRICING WITH NON-STANDARD RISKS (Room: R03) . . . 79
CO071: CONTRIBUTIONS IN BAYESIAN ECONOMETRICS (Room: R04) . . . 79
CO061: APPLIED NETWORK ANALYSIS IN EMPIRICAL FINANCE (Room: R06) . . . 80
CO660: THE MACROECONOMIC PROPAGATION OF SHOCKS (Room: R08) . . . 81
CO289: ADVANCES IN ROBUST ESTIMATION AND INFERENCE: THEORY AND APPLICATIONS II (Room: R19) . . . 81
CG336: CONTRIBUTIONS IN TIME SERIES AND FORECASTING (Room: R02) . . . 82
CG116: CONTRIBUTIONS IN STATISTICS AND ECONOMETRICS FOR THE COVID-19 PANDEMIC (Room: R07) . . . 82

Parallel Session J – CFE-CMStatistics (Sunday 20.12.2020 at 15:05 - 16:20) **84**
EO429: NONLINEAR METHODS IN FUNCTIONAL DATA ANALYSIS (Room: R11) . . . 84
EO708: RECENT DEVELOPMENTS IN MODEL-BASED CLUSTERING (Room: R12) . . . 84
EO213: RECENT DEVELOPMENTS IN IMAGING DATA ANALYSIS (Room: R13) . . . 85
EO359: RECENT DEVELOPMENT IN HIGH DIMENSIONAL METHODS (Room: R14) . . . 85
EO445: ADVANCES IN THE STATISTICAL ANALYSIS OF DEPENDENT NETWORK DATA (Room: R15) . . . 86
EO437: RECENT ADVANCES IN THEORY AND METHODS FOR SPATIOTEMPORAL MODELING (Room: R16) . . . 86
EO534: STATISTICAL ADVANCES ON MICROBIOME DATA ANALYSIS I (Room: R18) . . . 87
EO606: CAUSAL INFERENCE METHODS IN GENETIC STUDIES (Room: R20) . . . 87
EO636: ADVANCES IN CAUSAL SURVIVAL ANALYSIS (Room: R21) . . . 88
EO760: BAYESIAN DATA INTEGRATION OF COMPLEX OBJECTS (Room: R22) . . . 88
EO510: STATISTICAL LEARNING FOR DECISION-MAKING SYSTEMS (Room: R23) . . . 89
EO323: ALGORITHMIC FAIRNESS WITH STATISTICAL GUARANTEES (Room: R25) . . . 89
EC787: CONTRIBUTIONS IN STATISTICAL MODELLING (Room: R24) . . . 90
CO618: ADVANCES IN FINANCIAL ECONOMETRICS (Room: R02) . . . 90
CO397: QUANTITATIVE APPROACH TO HIGHER EDUCATION RESEARCH (Room: R04) . . . 91
CO045: MACHINE LEARNING TECHNIQUES IN MACROECONOMICS AND FINANCE (Room: R07) . . . 91
CC804: CONTRIBUTIONS IN FORECASTING I (Room: R06) . . . 92
CG030: CONTRIBUTIONS IN FINANCIAL ECONOMETRICS II (Room: R03) . . . 92

Parallel Session K – CFE-CMStatistics (Sunday 20.12.2020 at 16:30 - 18:10) **93**
EO339: RANDOM OBJECTS: REGRESSION, CLUSTERING AND CHANGE-POINTS (Room: R11) . . . 93
EO447: NEW RESEARCH DIRECTIONS IN FUNCTIONAL DATA ANALYSIS (Room: R12) . . . 93
EO596: CLIMATE EXTREMES AND DEPENDENCE MODELING (Room: R13) . . . 94
EO496: PUBLIC POLICY ANALYSIS AND MACHINE LEARNING II (Room: R14) . . . 95
EO239: RECENT ADVANCES IN MULTIPLE HYPOTHESES TESTING (Room: R15) . . . 95
EO464: SPATIAL STATISTICS (Room: R16) . . . 96
EO287: METHODS ON HIGH DIMENSIONAL STATISTICS (Room: R17) . . . 96
EO536: STATISTICAL ADVANCES ON MICROBIOME DATA ANALYSIS II (Room: R18) . . . 97
EO153: SKETCHING AND RELATED METHODS IN REGRESSION (Room: R19) . . . 97
EO103: ADVANCES IN CAUSAL INFERENCE (Room: R20) . . . 98
EO540: STATISTICS FOR WEARABLE DEVICE DATA (Room: R21) . . . 99
EO620: RECENT DEVELOPMENTS IN BAYESIAN METHODOLOGY (Room: R22) . . . 99
EO227: BAYESIAN APPLICATIONS IN BIOLOGICAL AND ENVIRONMENTAL SCIENCES (Room: R23) . . . 100
EO377: SAMPLING AND CORESETS FOR LARGE-SCALE DATA (Room: R24) . . . 100
EO317: STATISTICAL MODELS: RECENT DEVELOPMENTS I (Room: R25) . . . 101
CI023: RECENT ADVANCES IN TIME SERIES ECONOMETRICS (Room: R02) . . . 101
CO149: SENTOMETRICS (Room: R04) . . . 102
CO193: APPLIED ECONOMETRICS (Room: R06) . . . 103
CO309: SELF-FULFILLING PROPHECIES AND MACROECONOMIC BEHAVIOR (Room: R07) . . . 103
CO544: TOPICS IN MACROECONOMETRICS (Room: R08) . . . 104
CG613: CONTRIBUTIONS IN PORTFOLIO ANALYSIS AND ASSET PRICING (Room: R03) . . . 104

Parallel Session M – CFE-CMStatistics (Monday 21.12.2020 at 08:45 - 10:00) **106**
EO700: COPULA MODELS IN ECONOMETRICS (Room: R04) . . . 106
EO145: ADVANCES IN STATISTICAL MODELLING (Room: R05) . . . 106
EO466: CHALLENGES IN FUNCTIONAL DATA ANALYSIS AND VARYING COEFFICIENT MODELS (Room: R11) . . . 106
EO754: SOME RECENT STATISTICAL DEVELOPMENTS IN CLUSTERING (Room: R12) . . . 107
EO624: RECENT DEVELOPMENTS IN QUANTILE REGRESSION (Room: R14) . . . 107
EO722: ADVANCED TREE METHODS AND APPLICATIONS (Room: R15) . . . 107
EO550: MODERN APPROACHES TO DIRECTIONAL DATA ANALYSIS (Room: R17) . . . 108
EO630: CAUSAL SURVIVAL ANALYSIS (Room: R21) . . . 108
EO157: BAYESIAN MACHINE LEARNING (Room: R22) . . . 109
EO614: RECENT ADVANCES IN SCREENING DESIGNS (Room: R24) . . . 109
EG575: CONTRIBUTIONS IN BAYESIAN MODELLING (Room: R20) . . . 110
CO710: NONLINEAR, SEMI- AND NONPARAMETRIC PANEL DATA MODELING (Room: R02) . . . 110
CO119: TOPICS IN FINANCIAL ECONOMETRICS (Room: R03) . . . 110
CG094: CONTRIBUTIONS IN FINANCIAL MARKETS (Room: R06) . . . 111
CG020: CONTRIBUTIONS IN FORECASTING II (Room: R07) . . . 111

Parallel Session N – CFE-CMStatistics (Monday 21.12.2020 at 10:10 - 12:15) **113**
EO500: CLUSTERING OF MULTIVARIATE DEPENDENT DATA (Room: R12) . . . 113
EO179: ML-ECO: MACHINE LEARNING AND STATISTICAL TECHNIQUES FOR ECONOMICS (Room: R13) . . . 113
EO610: MODEL SPECIFICATION TESTS (Room: R14) . . . 114
EO225: TOPICS ON HIGH-DIMENSIONAL METHODOLOGY (Room: R15) . . . 115
EO073: TOPICS IN TIME SERIES (Room: R17) . . . 115
EO259: RECENT DEVELOPMENTS IN MULTIVARIATE DATA ANALYSIS (Room: R20) . . . 116
EO590: INFERENCE IN SURVIVAL MODELS (Room: R21) . . . 117
EO053: TRACK: BAYESIAN SEMI- AND NON-PARAMETRIC METHODS I (Room: R22) . . . 117
EO634: RECENT ADVANCES ON DESIGN OF EXPERIMENTS (Room: R24) . . . 118
EG070: STATISTICS FOR HILBERT SPACES II (Room: R11) . . . 119
EP002: POSTER SESSION I (Room: Poster 1) . . . 120
EP008: POSTER SESSION II (Room: Poster 2) . . . 121
CO097: REALIZED MEASURE ANALYSIS AND MODELING IN HIGH DIMENSION (Room: R04) . . . 121
CO716: BAYESIAN MACROECONOMETRICS (Room: R08) . . . 122
CC807: CONTRIBUTIONS IN TIME SERIES ECONOMETRICS (Room: R02) . . . 123
CG026: CONTRIBUTIONS IN FINANCIAL ECONOMETRICS III (Room: R03) . . . 124
CG256: CONTRIBUTIONS IN MODELLING, VOLATILITY AND ACCURACY (Room: R06) . . . 124
CG024: CONTRIBUTIONS IN CREDIT RISKS (Room: R07) . . . 125

Parallel Session P – CFE-CMStatistics (Monday 21.12.2020 at 14:15 - 15:55) **127**
EI009: ALGORITHMS AND HIGH STAKES POLICY DECISIONS (Room: R11) . . . 127
EO562: FUNCTIONAL DATA ANALYSIS (Room: R12) . . . 127
EO433: ADVANCES OF COMPLEX DATA ANALYSIS (Room: R13) . . . 128
EO564: ADVANCES IN STATISTICAL METHODS AND APPLICATION WITH DIGITAL DATA (Room: R14) . . . 128
EO085: RECENT DEVELOPMENTS ON ANALYSIS OF NETWORKS (Room: R15) . . . 129
EO441: RECENT ADVANCES IN ROBUST AND NONPARAMETRIC REGRESSION (Room: R16) . . . 129
EO201: STATISTICS FOR HIGH-DIMENSIONAL HIGH-FREQUENCY DATA (Room: R17) . . . 130
EO347: RECENT DEVELOPMENTS OF COMPETING RISK DATA (Room: R18) . . . 130
EO079: STATISTICS IN NEUROSCIENCE I (Room: R19) . . . 131
EO083: CAUSAL INFERENCE AND GRAPHICAL MODELS (Room: R20) . . . 132
EO654: RECENT STATISTICAL ADVANCES IN HIGH-DIMENSIONAL BIOMEDICAL APPLICATIONS (Room: R21) . . . 132
EO091: BAYESIAN MODEL COMPARISON (Room: R22) . . . 133
EO470: BAYESIAN CAUSAL MODELLING OF TREATMENT STRATEGIES (Room: R23) . . . 133
EO349: NEW DEVELOPMENTS IN SURVEY SAMPLING (Room: R24) . . . 134
EO387: BRANCHING PROCESSES: THEORY, COMPUTATION AND APPLICATIONS I (Room: R25) . . . 135
CO033: TIME SERIES ECONOMETRICS MEETS CROSS SECTIONAL HETEROGENEITY (Room: R02) . . . 135
CO161: SEMI- AND NONPARAMETRIC REGRESSION FOR TIME SERIES AND PANEL DATA (Room: R03) . . . 136
CO093: CLIMATE CHANGE ECONOMETRICS AND FINANCIAL MARKETS (Room: R06) . . . 136
CO474: APPLIED MACRO (Room: R08) . . . 137
CC810: CONTRIBUTIONS IN RISK ANALYSIS (Room: R05) . . . 137
CG022: CONTRIBUTIONS IN ECONOMETRIC INFERENCE (Room: R04) . . . 138
CG046: CONTRIBUTIONS IN MACHINE LEARNING TECHNIQUES IN MACROECONOMICS AND FINANCE (Room: R07) . . . 139

Parallel Session Q – CFE-CMStatistics (Monday 21.12.2020 at 16:05 - 17:45) **140**
EI013: STATISTICAL FOUNDATION FOR DATA SCIENCE (Room: R14) . . . 140
EO315: STATISTICAL MODELS: RECENT DEVELOPMENTS II (Room: R04) . . . 140
EO658: EFFICIENT NONPARAMETRIC METHODS FOR COMPLEX DATA (Room: R05) . . . 141
EO698: FUNCTIONAL AND COMPLEX DATA ANALYSIS (Room: R11) . . . 141
EO243: ADVANCES IN THE ANALYSIS OF FUNCTIONAL DATA AND FUNCTIONAL TIME SERIES (Room: R12) . . . 141
EO570: ADVANCES IN THE ANALYSIS OF LARGE AND COMPLEX DATA (Room: R13) . . . 142
EO516: COMPUTATIONAL ISSUES IN INFECTIOUS DISEASE EPIDEMIOLOGY (Room: R15) . . . 143
EO123: METHODS FOR MISSING DATA IN EHR-BASED STUDIES (Room: R16) . . . 143
EO435: ADVANCES AND CHALLENGES IN MICROBIOME DATA ANALYSES (Room: R17) . . . 144
EO728: STATISTICAL INFERENCE IN COMPLEX NETWORKS (Room: R18) . . . 144
EO137: STATISTICS IN NEUROSCIENCE II (Room: R19) . . . 145
EO185: NOVEL APPROACHES TO CAUSAL INFERENCE (Room: R20) . . . 146
EO736: RECENT ADVANCES IN BIOSTATISTICS (Room: R21) . . . 146
EO135: ADVANCES IN BAYESIAN METHODS AND APPLICATIONS (Room: R22) . . . 147
EO756: ADVANCES IN BAYESIAN REGRESSION MODELING (Room: R23) . . . 147
EO680: RECENT ADVANCEMENTS IN HIGH DIMENSIONAL STATISTICS (Room: R24) . . . 148
EO389: BRANCHING PROCESSES: THEORY, COMPUTATION AND APPLICATIONS II (Room: R25) . . . 149
CI027: APPLIED TIME SERIES (Room: R02) . . . 149
CO111: TIME SERIES ECONOMETRICS II (Room: R03) . . . 150
CO233: DEVELOPMENTS IN CRYPTOCURRENCY AND BLOCKCHAIN (Room: R07) . . . 150
CO229: INFLATION DYNAMICS AND COMMUNICATIONS (Room: R08) . . . 151
CC815: CONTRIBUTIONS IN FINANCIAL MODELLING AND APPLICATIONS (Room: R06) . . . 151

Saturday 19.12.2020 08:45 - 09:45 Room: R01 Chair: Tommaso Proietti Opening and keynote talk 1

Econometric methods for empirically modelling climate change

Speaker: **David Hendry, University of Oxford, United Kingdom** Jennifer Castle

Economic and climate time series share many commonalities. Both are subject to non-stationarities in the form of evolving stochastic trends and sudden, often unanticipated distributional shifts, and both face incomplete knowledge of the human behaviour generating the data (DGP). Consequently, the well-developed machinery for modelling economic time series can be fruitfully applied to observational climate time series. We discuss the model formulation and selection methodology for locating an unknown DGP nested within a large set of possible explanation while also allowing for dynamic feedbacks, outliers, shifts, and non-linearities. We focus on indicator saturation estimators to handle shifts. The approach is illustrated by investigating the causal role of CO2 in Ice Ages and the UKs highly non-stationary annual CO2 emissions over the last 150 years.

Saturday 19.12.2020 11:45 - 12:35 Room: R01 Chair: Steven Gilmour Keynote talk 2

Additive models: Smooth backfitting, high dimensions, random mixed forests

Speaker: **Enno Mammen, Heidelberg University, Germany**

The purpose is to report on recent developments in the study of nonparametric additive models where the expectations of response variables are modeled as the sum of nonparametric functions of covariables. We will discuss extensions and modifications of the model. We will consider the high-dimensional case where the number of additive components converges to infinity and sparsity assumptions are made. We will discuss random mixed forests, a modification of random forests designed for additive models, and nonparametric ANOVA type regression models.

Sunday 20.12.2020 18:20 - 19:10 Room: R01 Chair: Morten Nielsen Keynote talk 3

Conditional quantile coverage: An application to growth at risk

Speaker: **Valentina Corradi, University of Surrey, United Kingdom**

Tests for pairwise and multiple out-of-sample comparisons of parametric conditional quantile models are proposed. The tests rank the distance between actual and nominal conditional coverage w.r.t. the union of information sets across models, for a given loss function. Our approach operates uniformly over a compact set of quantile ranks, thereby assessing models' relative forecast ability across different quantile subsets. We derive the limiting distribution and establish the first-order validity of block bootstrap critical values. An empirical application to Growth-at-Risk (GaR) uncovers situations where a threshold quantile model improves over the standard linear quantile regression approach.

Monday 21.12.2020 13:15 - 14:05 Room: R01 Chair: Martin Wagner Keynote talk 4

Four Australian banks and the multivariate time-varying smooth transition correlation GARCH model

Speaker: **Timo Terasvirta, Aarhus University, Denmark** Anthony Hall, Annastiina Silvennoinen

Daily returns of four Australian banks, often called the Big Four, are modelled. For this purpose, we use a new multivariate volatility model that belongs to the family of conditional correlation GARCH models. The GARCH equations of this model contain a multiplicative deterministic component to describe long-run movements in volatility and, in addition, the correlations are deterministically time-varying. Parameters of the model are estimated jointly using maximum likelihood. Since the model is strongly nonlinear, we also present certain features of the estimation algorithm.

Monday 21.12.2020 17:55 - 18:55 Room: R01 Chair: Antonio Canale Keynote talk 5 and closing

Within, between, beyond: Methods for assessing variability in brain imaging

Speaker: **Michele Guindani, University of California, Irvine, United States**

An improved understanding of the heterogeneity of brain mechanisms is considered key for enabling the developments of targeted, precision, medicine interventions based on imaging features. We will describe a few Bayesian methods to characterize the heterogeneity typically observed both within- and between- subjects. First, we will describe models for multi-subject analysis that will identify population subgroups characterized by similar brain activity patterns, also by integrating available information on the subjects. Then, we will discuss methods to study changes in connectivity patterns over time, by combining analyses usually conducted through multiple steps into a single, unified, modeling framework that provides an accurate dynamic representation of brain processes. Finally, we will briefly address methods to characterize the association between a set of imaging as well as non-imaging predictors and an individual behavioral or clinical outcome. We will illustrate the performance of the methods in simulations and on real neuroimaging data.

Saturday 19.12.2020 09:55 - 11:35 Parallel Session B – CFE-CMStatistics

EO245 Room R11 RECENT ADVANCES IN FUNCTIONAL DATA ANALYSIS Chair: Yuko Araki

E1044: **Dynamic predictions via functional joint models for sparse longitudinal and time-to-event data**
Presenter: **Toshihiro Misumi**, Yokohama City University, Japan
Co-authors: Yuriko Takeda

Recently, joint modeling techniques of longitudinal and survival data have been frequently applied in the medical research area. In the framework of joint modeling, the model enables us to compute individual dynamic predictions of survival probabilities. The accuracy of prediction may be worse when longitudinal data are sparsely observed at irregular time-points since ordinal linear mixed-effects models are widely used in existing joint models. To overcome this issue, we propose a novel functional joint model for sparse longitudinal and survival data. We employ a reduced rank model for the longitudinal submodel to capture the trajectory of sparse longitudinal process accurately. A Cox proportional hazards model is applied to the survival submodel. Unknown parameters included in the model are estimated by a Bayesian approach. Some numerical examples are presented to demonstrate the effectiveness of our proposed modeling strategy.

E0686: **K-means clustering for sparsely sampled longitudinal data**
Presenter: **Michio Yamamoto**, Okayama University / RIKEN AIP, Japan
Co-authors: Yoshikazu Terada

In longitudinal data, the observations often occur at different time points for each subject. In such a case, the ordinary clustering algorithms, such as the K-means clustering, cannot be applied directly. One may apply a smoothing technique to get individual continuous trajectories, followed by finding groups among the trajectories using some clustering algorithm. However, this approach is not appropriate when data of each subject are observed at only a few time points. For sparsely sampled longitudinal data, we develop a new simple clustering algorithm, which can be considered a natural extension of the K-means. We show the consistency of the proposed estimator under mild regularity conditions. Moreover, we investigate the empirical performance of the proposed method through simulation studies and data applications.

E0719: **Locally differential private functional data analysis**
Presenter: **Masaaki Imaizumi**, The University of Tokyo, Japan

The regression problem with functional data under the local differential privacy model is investigated. A notion of *functional data* is a significant scheme for handling high-frequency data such as personal vital logs. Due to increasing data sources for such data, it is a promising problem to study functional data from the aspect of local differential privacy. We develop an LDP-FLR mechanism that provides an interactive local differentially private estimator for the functional linear regression problem. To develop the mechanism, we utilize an orthogonal basis representation for functional data and an iterative update with stochastic gradient descent. Thanks to the orthogonal representation, it can avoid an error and computational cost from perturbation for local differential privacy. We also develop an adaptive method for hyper-parameter selection and an unknown orthogonal basis. Consequently, we prove that an estimator by LDP-FLR for the regression model can attain the minimax optimality when functional data have rich information. In contrast, we find that error by the estimator can get larger if functional data are less informative due to an ill-posed structure of functional data regression. Experimental results support our theoretical claim.

EO159 Room R12 ADVANCES IN DIRECTIONAL STATISTICS Chair: Agnese Panzera

E1079: **Applications of errors-in-variables models to wind data**
Presenter: **Stefania Fensore**, University of Chieti-Pescara, Italy
Co-authors: Marco Di Marzio, Agnese Panzera, Charles C Taylor

Wind direction data are typically subject to biases of several degrees because measurements are affected by many factors. Then, wind directions can be regarded as circular data observed with error. We discuss some kernel-based methods for estimating circular densities when data are observed with error and propose them to tackle the prediction of wind direction distributions as an errors-in-variables problem.

E0354: **Sine-skewed toroidal distributions and their application in protein bioinformatics**
Presenter: **Christophe Ley**, Ghent University, Belgium
Co-authors: Jose Ameijeiras-Alonso

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 torus. The main drawback of most of these models is that the related densities are (pointwise) symmetric, even though the data usually present asymmetric patterns. This motivates the need to find a new way of constructing asymmetric toroidal distributions starting from a symmetric distribution. We tackle this problem by introducing the sine-skewed toroidal distributions. The general properties of the new models are derived. Based on the initial symmetric model, explicit expressions for the shape parameters are obtained, a simple algorithm for generating random numbers is provided, and asymptotic results for the maximum likelihood estimators are established. An important feature of our construction is that no extra normalizing constant needs to be calculated, leading to more flexible distributions without increasing the complexity of the models. The benefit of employing these new sine-skewed toroidal distributions is shown based on protein data, where, in general, the new models outperform their symmetric antecedents.

E0499: **Testing parametric regression models with circular response**
Presenter: **Andrea Meilan-Vila**, Universidade da Coruna, Spain
Co-authors: Mario Francisco-Fernandez, Rosa Crujeiras

Testing procedures for assessing a parametric regression model with a circular response and a $\mathbb{R}^d$-valued covariate are proposed and analyzed. The considered test statistics are based on a comparison between a (non-smoothed or smoothed) parametric fit under the null hypothesis and a nonparametric estimator of the circular regression function. More specifically, two different test statistics are designed. In the first one, a parametric estimator of the regression function under the null hypothesis is directly used. In contrast, in the second one, a smooth version of this estimator is employed. The null hypothesis that the regression function belongs to a certain parametric family is rejected if a suitable circular distance between both fits exceeds a certain threshold. Different bootstrap procedures to calibrate the tests in practice are presented. Finite sample performance of the tests in several scenarios is analyzed by simulations. An illustration with a real dataset is also provided.

E0520: **Elliptical symmetry models and robust estimation methods on spheres**
Presenter: **Janice Scealy**, Australian National University, Australia
Co-authors: Andrew Wood

First, a new distribution is proposed for analysing directional data that is a novel transformation of the von Mises-Fisher distribution. The new distribution has ellipse-like symmetry, as does the Kent distribution; however, unlike the Kent distribution the normalising constant in the new density is easy to compute an estimation of the shape parameters is straightforward. To accommodate outliers, the model also incorporates an additional shape parameter which controls the tail-weight of the distribution. Next, we define a more general semi-parametric elliptical symmetry model on the sphere and propose two new robust direction estimators, both of which are analogous to the affine-equivariant spatial median in Euclidean space. We calculate influence functions and show that the new direction estimators are standardised bias robust in the highly concentrated

case. To illustrate our new models and estimation methods, we analyse archaeomagnetic data and lava flow data from two recently compiled online geophysics databases.

EO662 Room R13 ESTIMATION METHODS FOR EXTREME EVENTS Chair: Simone Padoan

E0458: **Extreme expectile estimation for heavy-tailed time series**
Presenter: **Gilles Stupfler**, ENSAI - CREST, France
Co-authors: Simone Padoan

Expectiles are a least-squares analogue of quantiles which have lately received substantial attention in actuarial and financial risk management contexts. Unlike quantiles, expectiles define coherent risk measures and are determined by tail expectations rather than tail probabilities; unlike the Expected Shortfall, they define elicitable risk measures. This has motivated recent studies of the behaviour and estimation of extreme expectile-based risk measures. The case of stationary but weakly dependent observations has, however, been left largely untouched, even though correctly accounting for the uncertainty present in typical financial applications requires the consideration of dependent data. We investigate the estimation of, and construction of accurate confidence intervals for, extreme expectiles and expectile-based Marginal Expected Shortfall in a general beta-mixing context, containing the classes of ARMA, ARCH and GARCH models with heavy-tailed innovations that are of interest in financial applications. The methods are showcased in a numerical simulation study and on real financial data.

E0507: **Data imputation of large observations via Bayesian inference for multivariate extremes**
Presenter: **Isadora Antoniano-Villalobos**, Ca' Foscari University of Venice, Italy
Co-authors: Simone Padoan, Boris Beranger

Missing data is a known issue in statistics. In applications placing interest on large observations, usual data imputation methods may fail to reproduce the heavy tail behaviour of the quantities 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 effective data imputation of multiple components at adequately large levels, provided that the model used for the angular measure is flexible enough to capture complex dependence structures. A Bayesian nonparametric model based on constrained Bernstein polynomials ensures such flexibility while allowing for tractable inference. An additional advantage of this approach is the natural way in which uncertainty about the estimation is incorporated into the imputed values through the Bayesian paradigm.

E0525: **Estimation and uncertainty quantification for extreme quantile regions**
Presenter: **Boris Beranger**, University of New South Wales, Australia
Co-authors: Simone Padoan, Scott Sisson

Estimation of extreme quantile regions, spaces in which future extreme events can occur with a given low probability, even beyond the range of the observed data, is an important task in the analysis of extremes. Existing methods to estimate such regions are available, but do not provide any measures of estimation uncertainty. We develop univariate and bivariate schemes for estimating extreme quantile regions under the Bayesian paradigm that outperforms existing approaches and provides natural measures of quantile region estimate uncertainty. We examine the method's performance in controlled simulation studies and then explore its application to the analysis of multiple extreme pollutant occurrences in Milan, Italy.

E0746: **Consistency of Bayesian and empirical Bayesian inference on max-stable models**
Presenter: **Stefano Rizzelli**, EPFL, Switzerland
Co-authors: Simone Padoan

Predicting the extremes of multiple variables is important in many applied fields. The Bayesian approach provides a natural framework for statistical prediction. Although various Bayesian inferential procedures have been proposed in the literature of univariate extremes and some for multivariate extremes, the study of their asymptotic properties has been left largely untouched. We establish the strong posterior consistency of a semiparametric Bayesian inferential procedure for well-specified max-stable models of arbitrary dimension, whose margins can be short-, light- or heavy-tailed. We then extend our consistency results to the case where the data come from a distribution lying in a neighbourhood of a max-stable one, representing a more realistic inferential setting. In doing this, we define a new hybrid-Bayesian approach, where data-dependent priors are specified in an empirical Bayes fashion. The developed technical tools appear of independent interest, beyond the context of the extreme values, and can be adapted to other statistical methods affected by a model convergence bias.

EO638 Room R15 NON-PARAMETRIC ANALYSIS OF COMPLEX DATA Chair: Weichi Wu

E0457: **Shift identification in time varying regression quantiles**
Presenter: **Subhra Sankar Dhar**, IIT Kanpur, India
Co-authors: Weichi Wu

The purpose is to discuss whether time-varying quantile regression curves are the same up to the horizontal shift or not. The errors and covariates involved in the regression model are allowed to be locally stationary. We formalise this issue in a corresponding non-parametric hypothesis testing problem and develop an integrated-squared-norm based test (SIT), as well as a simultaneous confidence band (SCB) approach. The asymptotic properties of SIT and SCB under null and local alternatives are derived. We then propose valid wild bootstrap algorithms to implement SIT and SCB. Furthermore, the usefulness of the proposed methodology is illustrated for various simulated and real data related to social science and climate science.

E0834: **Adaptive estimation for evolutionary high-dimensional time series factor models and a test for static factor loadings**
Presenter: **Weichi Wu**, Tsinghua University, China
Co-authors: Zhou Zhou

The estimation and testing of a class of high-dimensional non-stationary time series factor models with evolutionary temporal dynamics are considered. In particular, the entries and the dimension of the factor loading matrix are allowed to vary with time while the factors and the idiosyncratic noise components are locally stationary. We propose an adaptive sieve estimator for the span of the varying loading matrix and the locally stationary factor processes. A uniformly consistent estimator of the effective number of factors is investigated via eigenanalysis of a non-negative definite time-varying matrix. A high-dimensional bootstrap-assisted test for the hypothesis of static factor loadings is proposed by comparing the kernels of the covariance matrices of the whole time series with their local counterparts. We examine our estimator and test via simulation studies and real data analysis.

E0847: **Spectral inference under complex temporal dynamics**
Presenter: **Jun Yang**, University of Oxford, United Kingdom

Unified theory and methodology are developed for the inference of evolutionary Fourier power spectra for a general class of locally stationary and possibly nonlinear processes. In particular, simultaneous confidence regions (SCR) with asymptotically correct coverage rates are constructed for the evolutionary spectral densities on a nearly optimally dense grid of the joint time-frequency domain. A simulation-based bootstrap method is proposed to implement the SCR. The SCR enables researchers and practitioners to visually evaluate the magnitude and pattern of the evolutionary

power spectra with an asymptotically accurate statistical guarantee. The SCR also serves as a unified tool for a wide range of statistical inference problems in time-frequency analysis ranging from tests for white noise, stationarity, and time-frequency separability to the validation for non-stationary linear models.

E1115: **Modeling heterogeneous networks in the presence of covariates**
Presenter: **Swati Chandna**, Birkbeck, University of London, United Kingdom

Many applications routinely observe covariates at node and dyad level in addition to pairwise interactions between the agents of interest. A nonparametric approach to modeling unlabeled networks is offered by the graphon function. Recently, there has been a growing interest on the problem of graphon estimation as well as its application to important problems such as bootstrapping networks, testing for equivalence of network distribution using subgraph counts, estimation of missing links. Existing histogram approximations to graphon function are not designed to estimate heterogeneity across the full network. We will discuss an approach to modeling heterogeneity in network data using covariates via the graphon model.

EO223 Room R16 PROJECTION PURSUIT Chair: Nicola Loperfido

E0551: **From multivariate to univariate: Projections of optimal portfolio selection problems**
Presenter: **Tomer Shushi**, Ben Gurion University of the Negev, Israel
Co-authors: Zinoviy Landsman, Udi Makov

The focus is on the problem of maximization of the functional of expected portfolio return and variance portfolio return in its most general form. We present an explicit closed-form solution of the optimal portfolio selection. This problem is closely related to expected utility maximization and two-moment decision models. We show that most known risk measures, such as mean-variance, expected shortfall, 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 by the authors concerning the maximization of a combination of expected portfolio return and a function of the variance of portfolio return. Our 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 belonged to the efficient frontier obtained for the mean-variance portfolio.

E0855: **Persistence via exact excursion time distributions**
Presenter: **Krzysztof Podgorski**, Lund University, Sweden
Co-authors: Georg Lindgren, Igor Rychlik

Finding the probability that a stochastic system stays in a certain region of its state space over a specified time – a long-standing problem in computational physics, applied and theoretical mathematics – is approached through the extended and multivariate Rice formula. In principle, it applies to any smooth multivariate process given that efficient numerical implementations of the high-dimensional integration are available. For Gaussian processes, the computations are effective and more precise than those based on the Rice series expansions and other approximations. It is shown that the approach can yield the explicit integral forms of a variety of excursion time distributional problems. It solves the two-step excursion dependence for a general stationary differentiable Gaussian process, in both theoretical and practical numerical sense. The solution is based on exact expressions for the probability density for one and two successive excursion lengths. The numerical routine RIND computes the densities using recent advances in scientific computing and is easily accessible for a general covariance function. Some analytical results are also offered that explain the effectiveness of the implemented method and points out how it can be utilized for non-Gaussian processes. The approach compares favorably with other methods. In particular, the approximation error is more controllable than in the independent and Markov interval approximations.

E0752: **The canonical kurtosis matrix**
Presenter: **Nicola Loperfido**, University of Urbino, Italy

The canonical kurtosis matrix of a p-dimensional random vector with finite fourth moments and positive definite covariance matrix is a $p \times p$ symmetric, positive definite matrix which conveniently summarizes the fourth standardized moments of the random vector itself. The applications of the canonical kurtosis matrix include cluster analysis, outlier detection, invariant coordinate selection, independent component analysis and projection pursuit. The main properties of the canonical kurtosis matrix are reviewed, and new ones are investigated, with special emphasis on sign-symmetry, skew-symmetry and exchangeability. The statistical applications of the canonical kurtosis matrix are illustrated with both real and simulated datasets.

E1193: **Clustering conditional higher moments with a model-based fuzzy procedure**
Presenter: **Massimiliano Giacalone**, University of Naples - Federico II, Italy
Co-authors: Roy Cerqueti, Raffaele Mattera

A new time series clustering procedure is considered, which allows for heteroskedasticity, non-normality and models non-linearity with a fuzzy approach. Specifically, considering a Generalized Autoregressive Score model, we propose to cluster time series according to their estimated conditional moments via the Autocorrelation-based fuzzy C-means (A-FCMd) algorithm. The usefulness of the proposed procedure is illustrated using several empirical applications with financial time series assuming both linear and nonlinear models specification and under several assumptions about time series density function. In the end, we provide a discussion on the possible use of projection pursuit with the aim of improving clustering performance.

EO077 Room R18 ADVANCES IN COPULA THEORY Chair: Elisa Perrone

E0972: **The Hellinger correlation**
Presenter: **Gery Geenens**, University of New South Wales, Australia

The defining properties of any valid measure of the dependence between two continuous random variables are revisited and complemented with two original ones, shown to imply other usual postulates. While other popular choices are proved to violate some of these requirements, a class of dependence measures satisfying all of them is identified. One particular measure, that we call Hellinger correlation, appears as a natural choice within that class due to both its theoretical and intuitive appeal. A simple and efficient nonparametric estimator for that quantity is proposed.

E0479: **A copula transformation in multivariate mixed discrete-continuous models**
Presenter: **Rosy Oh**, Ewha Womans University, Korea, South
Co-authors: Jae Youn Ahn, Sebastian Fuchs

Copulas allow flexible and simultaneous modeling of complicated dependence structures together with various marginal distributions. Especially if the density function can be represented as the product of the marginal density functions and the copula density function, this leads to both an intuitive interpretation of the conditional distribution and convenient estimation procedures. However, this is no longer the case for copula models with mixed discrete and continuous marginal distributions, because the corresponding density function cannot be decomposed so nicely. We introduce a copula transformation method that allows representing the density function of a distribution with mixed discrete and continuous marginals as the product of the marginal probability mass/density functions and the copula density function. With the proposed method, conditional

distributions can be described analytically, and the computational complexity in the estimation procedure can be reduced depending on the type of copula used.

E0504: **Model selection and goodness of fit tests for conditional copula models**
Presenter: **Jacco Wielaard**, University of Twente, Netherlands
Co-authors: Alexis Derumigny

Conditional copulas, also known as copula models with covariates, are models that describe the dependence between several random variables of interest, conditionally to some known explanatory variables. It is often assumed that these conditional copulas belong to a fixed parametric family, with a (conditional) parameter depending on the explanatory variables. We propose several goodness-of-fit tests for the assumption of good specification of a parametric conditional copula model, without any constraint on the conditional margins. Two such tests that use different bootstrap resampling procedures are compared in a simulation study. Finally, these tests are applied to a dataset of financial returns.

E0388: **On extremal problems for integrals with respect to a copula measure**
Presenter: **Claudio Ignazzi**, Universita del Salento, Lecce, Italy, Italy
Co-authors: Fabrizio Durante, Juan Fernandez Sanchez, Wolfgang Trutschnig

Extremal problems for functionals of type $\mu_C \mapsto \int_0^1 \int_0^1 F d\mu_C$ are considered, where μ_C is a copula measure and F is a Riemann integrable function on $[0,1]^2$ of a specific type. Such problems have been considered mainly in the study of limit points of two uniformly distributed sequences. Still, they are of potential interest also in the study of possible novel copula-based measures of association.

EO199 Room R19 COMPUTATIONAL AND THEORETICAL STATISTICS FOR STOCHASTIC PROCESSES Chair: Nakahiro Yoshida

E0959: **MCMC algorithms for posteriors on matrix spaces**
Presenter: **Kengo Kamatani**, ISM, Japan
Co-authors: Alexandros Beskos

Markov chain Monte Carlo (MCMC) algorithms are studied for target distributions defined on matrix spaces. Such an important sampling problem has yet to be analytically explored. We carry out a major step in covering this gap by developing the proper theoretical framework that allows for the identification of ergodicity properties of typical MCMC algorithms, relevant in such a context. Beyond the standard Random-Walk Metropolis (RWM) and preconditioned Crank–Nicolson (pCN), a novel algorithm, termed the 'Mixed' pCN (MpCN), is developed. RWM and pCN are shown not to be geometrically ergodic for an important class of matrix distributions with heavy tails. In contrast, MpCN has very good empirical performance within this class. Geometric ergodicity for MpCN is not fully proven, as some remaining drift conditions are quite challenging to obtain owing to the complexity of the state space. We do, however, make a lot of progress towards a proof, and show in detail the last steps left for future work. We illustrate the computational performance of the various algorithms through simulation studies, first for the trivial case of an Inverse-Wishart target, and then for a challenging model arising in financial statistics.

E0360: **LAD estimation of locally stable SDE**
Presenter: **Hiroki Masuda**, Kyushu University, Japan
Co-authors: Alexei Kulik

The goal is to prove the asymptotic (mixed) normality of the least absolute deviation (LAD) type estimator of locally stable SDE observed at high frequency, where the target drift coefficient may be nonlinear in both state variable and parameter. The proof essentially relies on the recently developed general representation result about small-time stable approximation for general locally stable processes. The result is a far-reaching extension of a previous study.

E0881: **Total variation distance between SDE driven by stable processes and their Euler scheme**
Presenter: **Arnaud Gloter**, Universite d Evry Val d Essonne, France
Co-authors: Emmanuelle Clement

The focus is on the rate of convergence to zero for the total variation distance between a class of stochastic differential equations and their Euler scheme. We consider $(X_{i/n})_{i=0,\dots,n}$, a discrete sampling of X solution of a stochastic differential equation driven by a pure jump Lvy process of alpha-Stable type, and $(\hat{X}_{i/n})_{i=0,\dots,n}$, the associated Euler scheme. We give an upper bound for the T.V. distance between the laws of $(X_{i/n})_{i=0,\dots,n}$ and $(\hat{X}_{i/n})_{i=0,\dots,n}$ as $n \to \infty$.

E0447: **State-dependency and Hawkes processes for order book modeling**
Presenter: **Ioane Muni Toke**, CentraleSupelec, France
Co-authors: Sfendourakis Emmanouil

The modeling of high-frequency occurrences of events in electronic limit order books has been for the past few years an active subject of research for both academic and practitioners, with potential applications in the understanding of market microstructure, execution problems, trading strategies development, market regulation, etc. Several contributions in this field use self-exciting Hawkes processes to account for the clustering of events. Recently, a state-dependent Hawkes process has been proposed to model high-frequency financial data, in which the excitation kernel is state-dependent. We investigate an alternative extension of Hawkes processes, in which state-dependency is added to the process intensity via a multiplicative term. More precisely, the multiplicative term is an exponential of a linear combination of observed state covariates, such as spread or the imbalance. From a practical point of view, this alternative definition allows using multiple exponential kernels without a cumbersome explosion of the parameter space dimension. It can be applied to all series of financial events, without having to track all-state modifications. We present empirical results for model selection as well as event type prediction for 30+ stocks traded on the French market in 2015.

EO538 Room R20 ANALYZING COMPOSITIONAL, DISTRIBUTIONAL AND RELATIVE ABUNDANCE DATA I Chair: Karel Hron

E0381: **Independent component analysis for compositional data**
Presenter: **Kamila Facevicova**, Palacky University Olomouc, Czech Republic
Co-authors: Christoph Muehlmann, Klaus Nordhausen, Alzbeta Gardlo

Compositional data represent a family of multivariate data whose (strictly positive) parts carry relative information about the respective structure primarily. Due to this specific nature of the data, its direct analysis using standard multivariate statistical methods is not appropriate since it can lead to spurious results. As a way out, either the methods need to be modified with respect to the log-ratio methodology or the compositional dataset has to be expressed in a proper system of real-valued coordinates. The focus is on the adaptation of independent component analysis to the case of a compositional dataset. Independent component analysis aims at finding statistically independent components in the data and beside the dimension reduction, it is also suitable to search for groups within the data as well as outlying observations. The performance of the proposed methodology will be demonstrated on a metabolomics dataset.

E0706: **Group comparison with count-based compositional data**
Presenter: **Jan Graffelman**, Universitat Politecnica de Catalunya, Spain
Co-authors: Juan Jose Egozcue, Vera Pawlowsky-Glahn

Compositional data consists of data vectors containing relative information on the parts of some whole. Such data is subject to a constraint, something that is particularly clear in the case of data vectors that consist of percentages summing up to 100%. The log-ratio transformation has been widely used to deal with compositional data, resulting in transformed data that is used as input for standard statistical procedures. Many compositional data sets are ultimately derived from counts that are expressed as fractions of a total. Asymptotically, log-ratio transformed compositions that are obtained from underlying multinomial counts follow the multivariate normal distribution, for which a theoretical covariance matrix can be obtained by using the delta method. This opens up the way to estimate the covariance matrix of the log-ratio coordinates in different ways, with either conventional sample-based estimators, or asymptotic theory inspired alternative estimators. Multivariate group comparisons by Hotelling's T^2 test or Wilks lambda rest, in the compositional setting, on the covariance matrix of the log-ratio coordinates. We evaluate with simulation studies which estimator for the covariance matrix of the log-ratio coordinates works best in different settings.

E0675: **Assessing the impact of covariates in compositional regression models**
Presenter: **Christine Thomas-Agnan**, Universite Toulouse, France

Regression models are considered which involve compositional vectors (i.e. carrying relative information) either as the dependent variable, as explanatory variables or on both sides of the regression equation. Measuring the marginal impacts of covariates in these models is not straightforward since a change in one component of a closed composition automatically affects the rest of the composition. A natural tool for this assessment is the concept of elasticity or semi-elasticity, depending on the model. Indeed the elasticity (respectively the semi-elasticity) is linked to the simplicial derivative of the expected value of the dependent variable with respect to the covariate in the case both the dependent and the explanatory variables are compositional (respectively in the case one of them is compositional). We present examples of derivations of such relationships. After recalling the formulas for evaluating these (semi-) elasticities as a function of parameters estimates, we illustrate their use on several examples. We contrast with alternative interpretations of these models. We describe several extensions for example to models including a total variable among the explanatory variables, with possibly several definitions of the total, and models involving a spatial dimension.

E0707: **Combinatorial regression model in abstract simplicial complexes**
Presenter: **Andrej Srakar**, University of Ljubljana, Slovenia

In the regression analysis of market share data, four main parametric type models are prevalent: multinomial logistic regression, attraction models of various types, Dirichlet covariance models, and compositional regression. We extend this arsenal of possibilities with a fifth type, a completely novel regression perspective labelled combinatorial regression, based on combining n-tuplets of sampling units into groups and treating them as compositions. The novel perspective, estimated using Bradley-Terry based maximum likelihood approach, allows an extensive number of perspectives in the analysis of, for example, triplets, quadruplets or quintuplets or using as a measure of disparity between the units (to construct regressors) different divergence measures. It also allows applications to very small datasets as the number of units in the new model can be expressed in terms of factorial products of units of the original sample. We provide the analysis of the new approach for triplets and quadruplets using Jensen-Shannon and generalized Jensen-Shannon divergence measures and provide the Gaussian asymptotic limits of the approach with exploring its properties also in a Monte Carlo simulation study. In a short application, we analyze sessile hard-substrate marine organisms image data from Italian coast areas, which allows exploring the new approach in relative abundance data setting.

EO480 Room R22 BAYESIAN COMPUTATION Chair: Gael Martin

E0279: **Multifidelity methods for approximate Bayesian computation**
Presenter: **Thomas Prescott**, University of Oxford, United Kingdom
Co-authors: Ruth Baker

A vital stage in the mathematical modeling of real-world dynamical systems is the calibration of a model's parameters to observed data. Likelihood-free parameter inference methods, such as approximate Bayesian computation (ABC), build Monte Carlo estimates by comparing the experimental data with large numbers of model simulations. However, the computational expense of generating these simulations forms a significant bottleneck in the practical application of such methods. Simulations of corresponding cheap, low-fidelity models can be used to reduce the computational expense of building these samples, at the cost of introducing additional variance to the resulting estimates. The variance costs and computational benefits can optimally balanced, to characterize the optimal choice of how often to simulate from cheap, low-fidelity models in place of expensive, high-fidelity models in Monte Carlo ABC algorithms. The resulting early accept/reject multifidelity ABC algorithm is shown to give improved performance over high-fidelity approaches in the context of both importance sampling and sequential Monte Carlo sampling.

E0282: **Approximate Bayesian computation through Gibbs like steps**
Presenter: **Gregoire Clarte**, Universita Paris Dauphine, France
Co-authors: Christian Robert, Robin Ryder, Julien Stoehr

Approximate Bayesian computation methods are useful for generative models with intractable likelihoods. These methods are however sensitive to the dimension of the parameter space, requiring exponentially increasing resources as this dimension grows. To tackle this difficulty, we explore a Gibbs version of the ABC approach that runs component-wise approximate Bayesian computation steps aimed at the corresponding conditional posterior distributions, and based on summary statistics of reduced dimensions. While lacking the standard justifications for the Gibbs sampler, the resulting Markov chain is shown to converge in distribution under some partial independence conditions. The associated stationary distribution can further be shown to be close to the true posterior distribution and some hierarchical versions of the proposed mechanism enjoy a closed form limiting distribution. Experiments also demonstrate the gain in efficiency brought by the Gibbs version over the standard solution.

E0917: **Variational approximation of factor stochastic volatility models**
Presenter: **Robert Kohn**, University of New South Wales, Australia
Co-authors: David Gunawan, David Nott

Estimation and prediction in high dimensional multivariate factor stochastic volatility models is an important and active research area because they allow a parsimonious representation of multivariate stochastic volatility. Such factor models are usually estimated by Markov chain Monte Carlo or particle methods, which are normally slow for high dimensional or long time series because of the large number of parameters and latent states involved. Fast batch and sequential variational methods are proposed to approximate the posterior distribution of the states and parameters in a factor stochastic volatility model. It also obtains one-step and multi-step ahead variational forecast distributions. The method is applied to simulated and real datasets and shown to produce good approximate inference and prediction compared to the latest particle Markov chain Monte Carlo approaches, but is much faster.

E0374: **Loss-based variational Bayes prediction**
Presenter: **Gael Martin**, Monash University, Australia

A new method is proposed for Bayesian prediction that caters for models with a large number of parameters and is robust to model misspecification. Given a class of high-dimensional (but parametric) predictive models, this new approach constructs a posterior predictive using a variational approximation to a loss-based, or Gibbs, posterior that is directly focused on predictive accuracy. The theoretical behaviour of the new prediction approach is analyzed, and a form of optimality demonstrated. Applications to Bayesian neural network models, autoregressive mixture models, and to the M4 forecasting competition, demonstrate that the approach provides more accurate results than various alternatives, including misspecified likelihood-based predictions.

EO183 Room R24 THE STEIN METHOD AND STATISTICS Chair: Robert Gaunt

E0636: **Functional approximations with Stein's method of exchangeable pairs**
Presenter: **Mikolaj Kasprzak**, University of Luxembourg, Luxembourg

Functional limit theorems are an important class of results describing convergence in distribution of Markov chains to diffusion processes. The convergence occurs in a space of functions (the Skorokhod space of cadlag paths) and after a proper rescaling of parameters, such that, in the limit, the chain is forced to make jumps infinitely often and its paths become continuous. Such results prove useful whenever the processes one wishes to model are discrete in nature, but it is more convenient to describe them using SDEs. This might be the case, for instance, if the conclusions one wishes to draw are easily obtainable using stochastic analysis or if one does not want changes in the local details to affect the model significantly. Researchers using functional approximations, for instance in population biology (where one may often switch from a discrete model to a continuous one after a proper rescaling and letting the population size go to infinity), are interested in measuring the error they make when doing so. Stein's method of exchangeable pairs turns out to be particularly useful in this case and provides powerful upper bounds on distances between the discrete and the limiting continuous processes. The theoretical setup for infinite-dimensional Stein's method, an abstract approximation theorem and concrete models it can be used for, will be considered. Among the examples, certain classes of U-processes and a graph-valued process will be discussed.

E0486: **To choose or not to choose a prior**
Presenter: **Fatemeh Ghaderinezhad**, Gent university, Belgium

The first challenging question in Bayesian statistics is how choosing the prior can affect the posterior distribution. How can the posteriors derived under different priors be similar as nowadays more and more data are collected? One of the newest and under development instruments to answer this question is Stein's method. This crafty method gives the lower and upper bounds to measure the distance of two posteriors derived under different priors (even improper priors), using the Wasserstein distance at fixed sample size. To this aim, we have proposed a methodology for tractable distributions with nested densities in one-dimensional settings. However, for practical purposes, the power of the Wasserstein distance idea has not at all been exploited so far. How can we quantify prior impact for any type of priors and any dimensions? To provide an answer to this question, we have introduced the Wasserstein Impact Measure (WIM) that relies on the numerical computation of the Wasserstein distances. It allows us to compare any two priors, thus making the WIM a fully usable alternative to the proposals from the literature.

E0873: **Minimum Stein discrepancy estimators**
Presenter: **Andrew Duncan**, Imperial College London, United Kingdom

When maximum likelihood estimation is infeasible, one often turns to score matching, contrastive divergence, or minimum probability flow to obtain tractable parameter estimates. We detail a unifying perspective of these techniques as minimum Stein discrepancy estimators and use this lens to design new diffusion kernel Stein discrepancy (DKSD) and diffusion score matching (DSM) estimators with complementary strengths. The main strength of this methodology is its flexibility, which allows one to design estimators with desirable properties for specific models at hand by carefully selecting a Stein discrepancy. We illustrate this advantage for several challenging problems for score matching, such as non-smooth, heavy-tailed or light-tailed densities.

E1093: **Measuring sample quality with diffusions**
Presenter: **Sebastian Vollmer**, University of Warwick & Alan Turing Institute, United Kingdom

Stein's method for measuring convergence to a continuous target distribution relies on an operator characterizing the target and Stein factor bounds on the solutions of an associated differential equation. While such operators and bounds are readily available for a diversity of univariate targets, few multivariate targets have been analyzed. We introduce a new class of characterizing operators based on It diffusions and develop explicit multivariate Stein factor bounds for any target with a fast-coupling It diffusion. As example applications, we develop computable and convergence-determining diffusion Stein discrepancies for log-concave, heavy-tailed and multimodal targets and use these quality measures to select the hyperparameters of biased Markov chain Monte Carlo (MCMC) samplers, compare random and deterministic quadrature rules and quantify bias-variance tradeoffs in approximate MCMC. Our results establish a near-linear relationship between diffusion Stein discrepancies and Wasserstein distances, improving upon past work even for strongly log-concave targets. The exposed relationship between Stein factors and Markov process coupling may be of independent interest.

EO594 Room R25 ASYMPTOTIC THEORY IN STATISTICS Chair: Zeng Li

E0396: **CLT for LSS of large dimensional Kendall rank correlation matrices and its applications**
Presenter: **Zeng Li**, Southern University of Science and Technology, China
Co-authors: Runze Li, Qinwen Wang

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 the Kendalls 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. The numerical comparisons demonstrate the robustness and superiority of our proposed test statistics under various mixed and heavy-tailed cases.

E0789: **Robust simultaneous registration and classification model**
Presenter: **Jian Qing Shi**, Southern Univesity of Science and Technology, China

An extended *t* process-based two-level model is discussed, which allows simultaneously classifying and aligning functional data and provides a robust approach against outliers (either disturbed curves or disturbed discrete observed points in some curves). We use a logistic regression model and a data registration model to align and model the data at the same time, and also allow the models to use both scalar and functional variables. The trained models are applied to classify new data via an iterative procedure. The performance of the model is illustrated on both simulated and real data.

E0790: **Mixing efficiency of trans-model MCMC algorithms in Bayesian phylogenetics**
Presenter: **Xiyun Jiao**, Southern University of Science and Technology, China
Co-authors: Ziheng Yang, Tomas Flouri

Bayesian models are commonly used for phylogenetic reconstruction. However, the corresponding Markov chain Monte Carlo (MCMC) methods, especially the trans-model ones, typically suffer from poor mixing properties. We have explored the relationship between the efficiency of trans-model MCMC and the proposal distributions it uses. Based on the findings, we have proposed some methods to design the proposal distributions so that the performance of trans-model MCMC algorithms can be enhanced. We used two toy examples to illustrate our findings and deployed both theoretical arguments and real data analyses in Phylogenetics to show the effects of our methods.

E1122: **Variable selection in distributed sparse regression under memory constraints**
Presenter: **Xuejun Jiang**, Southern University of Science and Technology, China

Variable selection is studied using the penalized likelihood method for distributed sparse regression with large sample size n under a limited memory constraint, where the memory of one machine can only store a subset of data. This is a much-needed research problem to be solved in the big data era. A naive divide-and-conquer method solving this problem is to split the whole data into N parts and run each part on one of N machines, aggregate the results from all machines via averaging, and finally obtain the selected variables. However, it tends to select more noise variables, and the false discovery rate may not be well controlled. We improve it by a special designed weighted average in aggregation. Theoretically, we establish asymptotic properties of the resulting estimators for the likelihood model with a diverging number of parameters. Under some regularity conditions, we establish oracle properties in the sense that our distributed estimator shares the same asymptotic efficiency as the estimator based on the full sample. A distributed penalized likelihood algorithm is proposed to refine the results in the context of general likelihoods. Furthermore, the proposed method is evaluated by simulations and a real example.

EC792 Room R14 CONTRIBUTIONS IN SPATIAL STATISTICS Chair: David Bolin

E0412: **Wild bootstrap for spatio-temporal data**
Presenter: **Daisuke Kurisu**, Tokyo Institute of Technology, Japan
Co-authors: Kengo Kato

A novel bootstrap method is introduced for spatial and spatio-temporal data. For this, we derive Gaussian approximations and propose the spatial wild bootstrap (SWB) for sample means of a random field observed at a finite number of locations in a sampling region. In particular, we give Gaussian and bootstrap approximations for probabilities that the normalized sample means of discretely observed random field hit hyperrectangles. Additionally, we show that our results can be applied to a wide class of multivariate Levy driven moving average random fields and discuss multiple temporal change point tests for spatio-temporal data.

E1064: **Structural information-complexity transfer in log-Gaussian Cox processes**
Presenter: **Adriana Medialdea**, Universidad de Granada, Spain
Co-authors: Jose Miguel Angulo, Jorge Mateu

Log-Gaussian Cox processes establish a flexible class of spatial point pattern models which allow the representation of a wide variety of dependency effects. Information and complexity measures constitute a useful tool for assessment of stochasticity and structural richness of a system. In this framework, based on a box-counting approach from simulation, diverse scenarios of log-Gaussian Cox processes are analyzed under different parameter configurations of the Matern covariance model, with the aim of characterizing the structural information-complexity transfer from the underlying intensity field to the resulting point pattern. Generalized entropy, divergence and complexity measures are computed, enabling both global and local comparisons of the distributions corresponding to the two phases involved in the realization of the processes. Sensitivity with respect to varying values of deformation parameters is also assessed. Ordinary and relative diversity indices provide a direct interpretation of the structural enrichment from the intensity field to the subsequently generated point pattern.

E1057: **Exceedance assessment based on functional thresholds and general spatial measures**
Presenter: **Jose Luis Romero**, University of Granada, Spain
Co-authors: Ana Esther Madrid, Jose Miguel Angulo

Spatial deformation is used in different application areas, such as image analysis or environmental studies, to represent certain forms of heterogeneity which can be explained by transformation of a reference stationary random field. An extremal analysis is often associated with the study of indicators related to geometrical characteristics of excursion sets defined by threshold exceedances. Derivation of results about threshold exceedance probabilities constitutes an open field in terms of the consideration of generalized forms of thresholds, among other aspects. In particular, the spatial deformation of a flow-type random field can be viewed as the deformation of a level-type state-rescaled random field, locally in terms of the Jacobian of the transformation. This case suggests the formalization of generalized scenarios involving functional thresholds and non-Lebesgue spatial measures.

E0436: **A cross-entropy method for spatial clustering of binary data**
Presenter: **Nishanthi Raveendran**, Macquarie University, Australia
Co-authors: Georgy Sofronov, David Bulger

Spatial data is very often heterogeneous, indicating that there may not be a unique statistical model describing the data. To overcome this issue, the data can be segmented into several homogeneous regions (or domains). Identifying such homogeneous domains and their boundaries is called spatial clustering (or segmentation) in spatial statistics. Spatial clustering is commonly used in many different fields, including epidemiology, criminology, and ecology. The focus is on spatially correlated binary data indicating the presence or absence of plant species observed over a two-dimensional lattice. Factoring out cluster label permutations yields a non-Euclidean model space. To address this, a hybrid of the Cross-Entropy method with a genetic algorithm is proposed. Voronoi tessellation is used to estimate the cluster centres and boundaries of such domains. The results illustrate that the proposed algorithm is effective in identifying homogeneous clusters in spatial binary data.

EC790 Room R21 CONTRIBUTIONS IN SURVIVAL ANALYSIS Chair: Jacobo de Una-Alvarez

E0303: **Modeling and analysis of data with confounding covariates and crossing of the hazard functions**
Presenter: **Ruta Levuliene**, Vilnius University, Lithuania
Co-authors: Vilijandas Bagdonavicius

Parametric models for the analysis of survival data with a possible crossing of hazard rates related to two treatment groups are introduced. A strategy for survival improvement through the application of time-varying treatment is discussed. Complete and right-censored data with possible confounding covariates are considered. Estimators of the crossing points are given. Chi-squared type goodness-of-fit tests for the considered models are given. Parametric tests for the absence of crossing of survival functions (and also for crossing of the hazard functions) hypothesis are proposed. For models with several baseline distributions, the power functions of the tests were investigated by simulation. Moreover, real and synthetic data analysis is presented.

E0743: **Accelerated failure time vs Cox proportional hazards mixture cure models**
Presenter: **Motahareh Parsa**, KU Leuven, Belgium
Co-authors: Ingrid Van Keilegom

A cure model is a useful model for analyzing failure time data in which some subjects experience the event of interest (the uncured subjects), and others don't (the cured subjects). We focus on mixture cure models, which rely on a model for the cure probability and a model for the survival function of the uncured subjects, conditional on a set of covariates. For the latter model, one often uses a Cox proportional hazards model. Despite the many advantages of this model, like its easy interpretation and the availability of software, the model suffers from some important drawbacks, like the cure threshold is the same for all values of the covariates. This might be unrealistic in situations, where covariates contain important information about the cure proportion and the event time of subjects under study. An alternative model is the accelerated failure time (AFT) mixture cure model. The cure threshold in this model depends on the covariates and leads, therefore to a more realistic and better fit of the data in many cases. We show that the AFT and the Cox model both fit the data well in the regions of sufficient follow-up, but differ drastically outside that region.

E0979: **Efron-Petrosian integrals for doubly truncated data with covariates: An asymptotic analysis**
Presenter: **Jacobo de Una-Alvarez**, University of Vigo, Spain
Co-authors: Ingrid Van Keilegom

In survival analysis, epidemiology and related fields, there exists an increasing interest in statistical methods for doubly truncated data. Double truncation appears with interval sampling and other sampling schemes and refers to situations in which the target variable is subject to two (left and right) random observation limits. Doubly truncated data require specific corrections for the observational bias, and this affects a variety of settings, including the estimation of marginal and multivariate distributions, regression problems, and multi-state models. Multivariate Efron-Petrosian integrals for doubly truncated data are introduced. These integrals naturally arise when the goal is the estimation of the mean of a general transformation which involves the doubly truncated variable and covariates. An asymptotic representation of the Efron-Petrosian integrals as a sum of iid terms is derived and, from this, consistency and distributional convergence are established. As a by-product, uniform iid representations for the marginal nonparametric maximum likelihood estimator and its corresponding weighting process are provided. Applications to correlation analysis, regression, and competing risks models are presented. A simulation study is reported too.

E0510: **On semiparametric modelling, estimation and inference for survival data subject to dependent censoring**
Presenter: **Negera Wakgari Deresa**, KU Leuven, Belgium
Co-authors: Ingrid Van Keilegom

When modelling survival data, it is common to assume that the survival time T is conditionally independent of the censoring time C given a set of covariates. However, there are numerous situations in which this assumption is not realistic. The goal is therefore to develop a semiparametric normal transformation model, which assumes that after a proper nonparametric monotone transformation, the vector (T,C) follows a linear model, and the vector of errors in this bivariate linear model follows a standard bivariate normal distribution with a possibly non-diagonal covariance matrix. We show that this semiparametric model is identified, and propose estimators of the nonparametric transformation, the regression coefficients, and the correlation between the error terms. It is shown that the estimators of the model parameters and the transformation are consistent and asymptotically normal. We also assess the finite sample performance of the proposed method by comparing it with an estimation method under a fully parametric model. Finally, the method is illustrated using data from the AIDS Clinical Trial Group 175 study.

CO037 Room R02 TOPICS IN TIME SERIES ECONOMETRICS Chair: Martin Wagner

C0472: **Testing linear cointegration against smooth transition cointegration**
Presenter: **Martin Wagner**, University of Klagenfurt, Austria
Co-authors: Oliver Stypka

Tests are developed for the null hypothesis of linear cointegration against the alternative of smooth transition cointegration. The test statistics are based on the fully modified or integrated modified OLS estimators suitably modified to Taylor approximations of smooth transition functions. This necessitates the adaptation of the estimation mentioned above approaches to models, including cross-products of integrated regressors. As transition variables, we consider integrated variables and time. For the integrated modified OLS based test, we also develop fixed-b inference. The properties of the tests are evaluated with a simulation study and compared to other test proposed by previously.

C0566: **Identifiability and estimation of possibly non-invertible SVARMA models: A new parametrisation**
Presenter: **Bernd Funovits**, University of Helsinki, Finland

The focus is on parameterisation, identifiability, and maximum likelihood (ML) estimation of possibly non-invertible structural vector autoregressive moving average (SVARMA) models driven by independent and non-Gaussian shocks. We introduce a new parameterisation of the MA polynomial matrix based on the Wiener-Hopf factorisation (WHF) and show that the model is identified in this parametrisation for a generic set in the parameter space (when certain just-identifying restrictions are imposed). In particular, this parmetrization allows for MA zeros at zero, which can be interpreted as informational delays. Typically imposed identifying restrictions on the shock transmission matrix as well as on the determinantal root location are made testable. Furthermore, we provide low-level conditions for asymptotic normality of the ML estimator and analytic expressions for the score and the information matrix. As an application, we analyze non-invertibility (and in particular delays) in a standard macro-econometric model. This and further analyses are implemented in a well documented R-package.

C0628: **Cointegrating polynomial regressions with integrated regressors with drift: Fully modified OLS estimation and inference**
Presenter: **Karsten Reichold**, University of Klagenfurt, Austria
Co-authors: Martin Wagner

Although many macroeconomic variables are well described as integrated processes with drift, the cointegrating polynomial regression (CPR) literature lacks a complete analysis of the implications of the presence of integrated regressors with drift on estimation and inference. We address this issue by developing a fully modified (FM-)OLS estimator for CPRs with integrated regressors with potentially (unknown) non-zero drift. In case the deterministic regressors and the powers of the stochastic regressor share at least one identical power of time, the ensuing asymptotic multicollinearity needs to be addressed to develop asymptotic theory. Although the limiting distribution of the FM-OLS estimator is not invariant to the presence of an unknown and non-zero drift, it allows for standard asymptotic inference for Wald-type hypotheses and common specification tests. However, the limiting distributions of widely used (non-)cointegration test statistics depend on the presence of a non-zero drift. Corresponding critical values are provided. Simulation results show that the developed estimator and tests based upon it perform well in finite samples and also highlight that not taking a non-zero drift into account has substantial detrimental effects on estimator and test performance. In particular, the test for the null hypothesis of cointegration exhibits serious over-rejections.

C0821: **Monitoring structural break in error correction models**
Presenter: **Leopold Soegner**, Institute for Advanced Studies, Austria
Co-authors: Martin Wagner

Consistent monitoring procedures are developed to detect structural changes in a Johansen-type error correction model. In particular, we consider breaks where the cointegration rank remains constant. Furthermore, we investigate structural breaks in the parameters affecting the autoregressive and deterministic terms. We develop Lagrange multiplier tests allowing to monitor these kinds of breaks.

CO065 Room R03 HIGH FREQUENCY DATA IN ECONOMICS AND FINANCE Chair: Leone Leonida

C0304: **An inside look to the spectrum of realized betas**
Presenter: **Malvina Marchese**, Cass Business School, United Kingdom
Co-authors: Rodrigo Hizmeri, Marwan Izzeldin

The impact of the generally ignored log-price drift in estimating realized betas is investigated. The presence of non-negligible drift results in poor measures of the realized betas, mainly because the non-negligible drift negatively impact the realized (co)variance measures. Our Monte Carlo experiment confirms the dramatic impact of the drift-term not only in the generic realized beta, but also in the continuous and jumps betas. The main distortions occurred via three channels: i) high levels of the drift-term in the stock and/or the index; ii) when the drift-term of the stock and the index differ greatly; iii) as the sampling frequency decreases. We propose an alternative approach that mitigates the impact of the temporary

non-negligible drift, yielding more accurate estimates of the realized betas and their risk-premia. We illustrate the usefulness of our new approach using all the Dow Jones constituents for a period of 17 years.

C0338: **High-frequency options data in estimating time-varying risk-neutral densities using kernel-type estimators**
Presenter: **Ana Monteiro**, University of Coimbra, Portugal
Co-authors: Antonio Santos

High-frequency data allow developing new models and methods to analyze financial risk. The extraction of risk-neutral densities through option prices has been extensively treated in the literature. The best methods to extract such densities are still not fully established, and only recently, high-frequency data constituted an innovative factor. The risk-neutral density is associated with the second derivative of the option pricing function. We propose a nonparametric-based technique to estimate risk-neutral densities in a static environment and allowing extensions to dynamic ones. We define a criterion function used in nonparametric regression that includes calls, puts, and weights in a constrained optimization problem. The problem's constraints represent the no-arbitrage ones, which results in a large scale convex quadratic programming problem. Two of the main elements influencing option pricing function is the strike and the time to maturity. The former allows the definition of the risk-neutral density (second derivative). The latter allows a dynamic view of the risk-neutral density, which, compared with the "physical" density, permits a better understanding of economic-agents risk-aversion evolution. We developed an optimization problem inserted in a nonparametric estimation framework that allows through high-frequency option price data to estimate time-varying risk-neutral densities. We tested the methods using options contracts on the S&P500 futures.

C0347: **Factor models with downside risk**
Presenter: **Daniele Massacci**, Kings College London, United Kingdom
Co-authors: Lucio Sarno, Lorenzo Trapani

We propose a conditional model of asset returns in the presence of common factors and downside risk. Specifically, we generalize existing latent factor models in three ways: we allow for downside risk via a threshold specification which allows for the estimation of the (usually set a priori) 'disappointment event'; we permit different factor structures (and number of factors) in different regimes; we show how to recover the observable factors risk premia from the estimated latent ones in different regimes. The usefulness of this generalized model is illustrated through three applications to low-dimensional, medium-sized, and large cross-sections of asset returns.

C0659: **BrExit or BritaIn: Is the UK more attractive to supervisors? An analysis of wage premium to supervision across the EU**
Presenter: **Leone Leonida**, King's College London, United Kingdom
Co-authors: Antonio Giangreco, Sergio Scicchitano, Marco Biagetti

The wage premium to supervision (WPS) is defined as the extra wage that supervisors earn relative to their subordinates, and estimate it at different quantiles of wage distribution for 26 European economies, comparatively focusing on the UK. We find that, by compensating supervisory positions according to the wage, the WPS increases inequality in most of the economies. Further, over 10% of the WPS depends on the context. Our results support the hypothesis that, in regard to the WPS, the UK is more rewarding than the other economies. We discuss implications for immigration and policymakers in relation to the post-Brexit process.

CO063 Room R04 ADVANCES IN ROBUST ESTIMATION AND INFERENCE: THEORY AND APPLICATIONS I Chair: Artem Prokhorov

C0781: **Robust confidence sets for moment-based models**
Presenter: **Anton Skrobotov**, Russian Presidential Academy of National Economy and Public Administration and SPBU, Russia
Co-authors: Artem Prokhorov

A new approach is proposed to construct conservative confidence intervals when moment conditions define a large set of models with varying degrees of plausibility as measured by commonly used statistics. The statistics reflect orthogonality, relative efficiency and non-redundancy of the moments for a given model and allow for a multi-criterion search over all available moment combinations. Using the most plausible model for inference requires adjustments to the usual confidence intervals. We show how to obtain the adjustments, and we illustrate that they work very well in Monte Carlo simulations.

C1111: **Double-lasso estimation of Heckman's sample selection model**
Presenter: **Masayuki Hirukawa**, Ryukoku University, Japan
Co-authors: Di Liu, Irina Murtazashvili, Artem Prokhorov

Sample selection models that contain high-dimensional covariates in both the main and selection equations are investigated. The particular focus is on estimation and inference of a low-dimensional parameter of interest in the main equation. Taking econometric practices into account, we maintain the assumption of bivariate normality on the error terms in two equations and adopt the two-step estimation approach. A double-selection procedure based on l_1-penalized regression models is applied in each step. In the first step, we estimate the selection equation by tailoring the double-selection procedure for generalized linear models to the Probit model with many covariates. After obtaining an estimate of the inverse Mills ratio, we proceed to the second step, in which a double-selection procedure for linear regression models with many covariates is employed for the main equation. It is demonstrated that under the sparsity assumption, the estimator of the low-dimensional parameter in the main equation is $n^(1/2)$-consistent and asymptotically normal, where n is the sample size. Monte Carlo simulations confirm attractive properties of the estimator, and an empirical application is considered.

C1150: **Exchange rate fluctuations and bank cost efficiency: Evidence from an emerging economy**
Presenter: **Mikhail Mamonov**, CERGE-EI, Czech Republic
Co-authors: Artem Prokhorov, Christopher Parmeter

Similarly to advanced countries, emerging economies may have substantial variation in banks' involvement in foreign currency operations. However, as opposed to advanced countries, these economies may exhibit much higher volatility of nominal exchange rates. In these circumstances, the banks bear additional costs due to revaluations of their international operations, especially during the periods of systemic depreciation of exchange rates. Such revaluations are often disconnected from the banks' productivity and cost-efficiency. For Russian banks between 2004 and 2020, we document that the revaluations heavily distort the cost-efficiency rankings. This conceals the actual distribution of cost inefficiency where banks with mid-range efficiency scores change their rankings in surprising ways after we account for the revaluations. Highly efficient and highly inefficiency banks preserve their rankings, that is exhibit tail-dependence. These results are robust to such bank characteristics as ownership type and the size of total assets, and they hold during crisis and non-crisis periods. We discuss how revaluations relate to market power and risk aversion, and we find that the Demsetz efficient structure hypothesis holds only when the revaluations are accounted for. We conjecture such results apply to emerging markets more generally and offer several policy recommendations.

C1026: **Art price determinants: Author, artwork, and auction features**
Presenter: **Oleg Kuldyshev**, Saint Petersburg State University, Russia
Co-authors: Alexander Semenov, Dmitry Grigoriev, Valeria Kolycheva

Recent evidence suggests that, usually, the work of an experienced artist is more expensive than the work of a beginner, the work of a man is more expensive than the work of a woman, a painting is more expensive than graphics, a landscape is more expensive than a portrait, and the most expensive auctions are in the evening. However, what if we contrast the influence of the author's age and sex on the price with the influence of

the works material and technique? We identified a few dozen features and divided them into author, artwork, and auction characteristics. For the author characteristics, we used demographic features, such as age, sex, and nationality, as well as biographical features, such as education and migration. The artwork characteristics included size, material, technique, and provenance. The auction characteristics included the place and time of sale and the transaction currency. By applying hedonistic regression on a dataset of more than 14,000 works by the most expensive authors, we will determine which characteristics (i.e., author, artwork, or auction) have the greatest influence on the price and what features (e.g., nationality, artwork size, provenance, city, date of sale) are the most important.

CO133 Room R06 REGULARIZATION AND NETWORK APPROACHES IN FINANCIAL APPLICATIONS Chair: Gabriele Torri

C0573: **Networks: Sparse penalized estimation methods using elastic-net penalty**
Presenter: **Davide Bernardini**, University of Trento, Italy
Co-authors: Emanuele Taufer, Sandra Paterlini

In the context of Gaussian Markov networks, three estimators for graphical models with an elastic-net penalty are developed and tested. The goal is to estimate the sparse precision matrix from which to retrieve both the underlying conditional dependence graph and partial correlation graph. The first estimator relies on using conditional penalized regressions to estimate the precision matrix. In contrast, the second approach is based on direct penalization of the precision matrix in the likelihood function. Finally, the third estimator relies on a 2-stages procedure that estimates the edge set firstly and then the precision matrix. Through simulations, we investigate the performances of the proposed methods on a large set of well-known network structures. We show how a 2-steps procedure can improve the estimate both of the sparsity pattern in the graph and the edges' weights when we are interested in reconstructing a partial correlation graph.

C0631: **Tail dependence and ESG scores: An empirical investigation**
Presenter: **Karoline Bax**, University of Trento, Italy
Co-authors: Ozge Sahin, Claudia Czado, Sandra Paterlini

Growths in environmental, social, and governance (ESG) trading activity has reinforced the volatility in the global financial market. Over the last two decades, increases in market globalization have intensified the asymmetric dependence among international equity markets and a new paradigm of investment has altered the landscape for financial practitioners. Consequently, asset managers and regulators called for more diligent risk management and sparked a search for additional risk factors. This research aims to question whether ESG scores can be used as tail-risk measures and aid in financial risk assessments. It does so, by analyzing the tail dependence structure of companies with a range of ESG scores using high-dimensional copula modeling and Value-at-Risk (VaR) calculations. Empirical findings on real-world data will be discussed.

C1039: **Penalized enhanced portfolio replication with asymmetric deviation measures**
Presenter: **Gabriele Torri**, University of Bergamo, Italy
Co-authors: Rosella Giacometti, Sandra Paterlini

The problem of enhanced portfolio replication is addressed. A strategy is proposed based on the minimization of novel risk deviation measures based on expectiles. Such risk measures allow accounting asymmetrically for the differences between the portfolio and the benchmark, favouring positive deviations compared to negative ones. The model nests the minimum TEV replication approach as one special case. Ridge and elastic-net regularization penalties are added to the model to control estimation error better and improve the out-of-sample performances. Simulations and real-world analyses on multiple datasets allow us to discuss the pros and cons of the different methods.

C1069: **Efficient methods for high-dimensional robust variable selection**
Presenter: **Wojciech Rejchel**, University of Warsaw, Poland
Co-authors: Malgorzata Bogdan, Konrad Furmanczyk

Variable selection is a fundamental challenge, if one works with large-scale data sets, that the number of predictors significantly exceeds the number of observations. In many practical problems (from genetics or biology) finding a small set of significant predictors is as important as accurate estimation or prediction. We investigate the variable selection problem in the single index model $Y = g(\beta' X, \varepsilon)$, where Y is a response variable, X is a vector of predictors, β is the true parameter, and ε is a random error. We make no assumptions on the distribution of errors, the existence of their moments etc. Moreover, g is an unknown function. We propose a computationally fast variable selection procedure, which is based on standard Lasso with response variables replaced by their ranks. If response variables are binary, our approach is even simpler: we treat their class labels as they were numbers and apply standard Lasso. Since our approaches lead to misspecified models, we start with establishing the relation between the true parameter β and parameters, which we estimate. Then we present theoretical and numerical results describing variable selection properties of the methods.

CO295 Room R07 ADVANCES IN CREDIT RISK MODELLING Chair: Anthony Bellotti

C0248: **Opening the black box: Deep quantile neural networks for loss given default prediction**
Presenter: **Maximilian Nagl**, University of Regensburg, Germany
Co-authors: Ralf Kellner, Maximilian Nagl, Daniel Roesch

A flexible combination of quantile regression and neural networks for Loss Given Default prediction is proposed. The results show a superior performance compared to linear quantile regression. This may be caused by non-linear behaviour in higher quantiles, especially in Europe. By using a novel feature importance measure, we quantify the importance and direction of every input variable. This makes neural networks as interpretable as linear models. Moreover, we show that the macroeconomy is up to two times more important in USA than Europe and increasing in quantiles. The macroeconomy is most important in the US, whereas in Europe collateralization is essential.

C0517: **A new Twitter based credit rating model methodology**
Presenter: **Leonie Tabea Goldmann**, University of Edinburgh, United Kingdom
Co-authors: Raffaella Calabrese, Jonathan Crook

A model is proposed to predict corporate credit ratings using tweets about companies and tweets from the companies themselves. We make three contributions to knowledge. First, we relate tweets from the companies and tweets about the companies to corporate credit ratings. Second, we create different Twitter predictors and compare their performance. More specifically, we compare the performance of the tweet frequency, sentiment scores that have been calculated using different lexicon-based approaches and n-grams that have been created using differential language analysis. Third, we analyze the predictive power of tweets within the four largest industry sections and compare the differences. We analyze data relating to NASDAQ or NYSE listed companies over 2011-2019. We find including information from tweets gives a better predictive performance compared to models that omit them.

C0521: **A study of credit risk model robustness through a crisis**
Presenter: **Anthony Bellotti**, University of Nottingham Ningbo China, China

The COVID-19 crisis has led many organizations to worry about how their predictive models will perform during the changing social and economic environment. Some financial institutions have reverted to simpler models or human judgement, avoiding more complex models that they consider less robust to change. In light of the uncertainty, others have applied conservative decision making such as lowing credit limits. In an attempt to quantify the effects of dramatic change in population caused by crisis on credit risk models, we conduct an empirical study based on models built

before and during the last financial crisis in 2009, using the Freddie Mac loan-level data for mortgage originations from 2004 to 2016. We compare credit risk models using logistic regression with more complex models, such as random forests and artificial neural networks, to determine how they perform during and after the financial crisis.

C0843: **Joint model of longitudinal and spatiotemporal survival data**
Presenter: **Victor Medina-Olivares**, University of Edinburgh, United Kingdom
Co-authors: Raffaella Calabrese, Jonathan Crook, Finn Lindgren

In credit risk analysis, it is generally of interest to model the time-to-event (survival) of a borrower according to two types of covariates: time-fixed and time-varying. When the latter presents possible endogeneity, usually seen in this context, it is preferable to incorporate them in a joint modelling approach that considers the mutual evolution of survival time and the endogenous covariates, rather than treat them separately. Moreover, it is increasingly common to incorporate geographical information about the borrower into the databases, giving way to models that also account for spatial clustering and its variation in time. A Bayesian hierarchical joint model of longitudinal and discrete survival data considering spatiotemporal frailties is proposed. This approach captures the survival effect due to the evolution of the unobserved heterogeneity among subjects located in the same region.

CO219 Room R17 STATISTICAL LEARNING OF HIGH DIMENSIONAL DATA Chair: Chengchun Shi

C0178: **Using statistical learning to model the monetary policy transmission mechanism**
Presenter: **Petre Caraiani**, Bucharest University of Economic Studies; Institute for Economic Forecasting, Romania

The transmission mechanism of monetary policy is studied using a large firm-level dataset on US firms. Drawing on modern machine learning techniques, the aim is to uncover the underlying transmission mechanism of monetary policy shocks in a high dimensional data environment. The research aims at uncovering New channels of monetary policy are uncovered, established ones are confirmed or discarded by relying on adapted statistical learning techniques.

C0200: **Does the Markov decision process fit the data: Testing for the Markov property in sequential decision making**
Presenter: **Chengchun Shi**, LSE, United Kingdom
Co-authors: Runzhe Wan, Rui Song, Wenbin Lu, Ling Leng

The Markov assumption (MA) is fundamental to the empirical validity of reinforcement learning. We propose a novel Forward-Backward Learning procedure to test MA in sequential decision making. The proposed test does not assume any parametric form on the joint distribution of the observed data and plays an important role for identifying the optimal policy in high-order Markov decision processes and partially observable MDPs. We apply our test to both synthetic datasets and a real data example from mobile health studies to illustrate its usefulness.

C0223: **Robust mean and eigenvalues regularized covariance matrix estimation**
Presenter: **Wenyu Cheng**, London School of Economics and Political Science, United Kingdom
Co-authors: Clifford Lam

Covariance matrix is a common tool for summarising linear relationships between variables. The topic is particularly of interest in the high-dimensional setting, where the classical sample covariance estimator is no longer optimal. Non-parametric eigenvalue shrinkage covariance estimator (NERCOME) offers one solution by shrinking extreme eigenvalues non-linearly. It makes no structural assumptions on the covariance matrix and guarantees a positive definite outcome. However, its good performance relies on the availability of twelve moments, which is hard to verify and often unsatisfied in the real world. The unpredictable presence of few outliers from highly asymmetric or fat-tailed distributions could significantly jeopardise its performance. We draw on recent developments from the robust statistics literature. Specifically, we focus on generalising Catoni loss function to alleviate the impacts of extreme observations. The improved influence function, now requiring the existence of just over one moment, produces narrower bounds on estimated means. Incorporating these findings, the robust NERCOME behaves consistently across different distributional settings, while maintaining overall estimation efficiency and other desirable properties as in NERCOME. We challenge the robust NERCOME with highly skewed and leptokurtic scenarios through extensive simulation studies. Applications in financial data are provided in the end.

C0944: **Manifold structure in graph embeddings**
Presenter: **Patrick Rubin-delanchy**, University of Bristol, United Kingdom

Statistical analysis of a graph often starts with embedding, the process of representing its nodes as points in space. How to choose the embedding dimension is a nuanced decision in practice, but in theory, a notion of true dimension is often available. In spectral embedding, this dimension may be very high. However, it is shown that existing random graph models, including graphon and other latent position models, predict the data should live near a much lower dimensional set. One may therefore circumvent the curse of dimensionality by employing methods which exploit hidden manifold structure.

CC809 Room R08 CONTRIBUTIONS IN MONETARY POLICY Chair: Davide Romelli

C0537: **Market-based long-term inflation expectations in Japan: A refinement on breakeven inflation rates**
Presenter: **Kazuhiro Hiraki**, Bank of Japan, Japan
Co-authors: Wataru Hirata

In Japan, the breakeven inflation rate (BEI), commonly used as a proxy for market-based long-term inflation expectations, has evolved lower than survey-based measures of long-term inflation expectations. The literature has pointed to three factors, other than long-term inflation expectations, that act as drivers of long-term BEI rates: (i) the deflation protection option premium of inflation-linked bonds, (ii) the liquidity premium of the bonds, and (iii) the spread between nominal and real term premia (the term premium spread). An affine term structure model is estimated to decompose Japan's BEI into long-term inflation expectations and these three other driving factors. The empirical results show that the deflation protection option premium for Japan's Inflation-Indexed Bonds (JGBi) has pushed the BEI up. In contrast, the liquidity premium of JGBi and the term premium spread have pulled it down, all having non-negligible contributions to developments in the BEI. This indicates that the evolution of Japan's BEI has been driven by these three factors as well as by the long-term inflation expectations of market participants. Consequently, the estimated long-term inflation expectations have evolved higher than the BEI throughout almost the entire estimation period.

C0443: **Monetary policy uncertainty and firm dynamics**
Presenter: **Stefano Fasani**, Queen Mary University of London, United Kingdom
Co-authors: Haroon Mumtaz, Lorenza Rossi

A FAVAR model with external instruments is used to show that monetary policy uncertainty shocks are recessionary and are associated with an increase in the exit of firms and a decrease in entry and in the stock price with total factor productivity rising in the medium run. To explain this result, we build a medium-scale DSGE model featuring firm heterogeneity and endogenous firm entry and exit. These features are crucial in matching empirical responses. Versions of the model with constant firms or exogenous firms exit are unable to re-produce the FAVAR response of firms entry and exit and suggest a much smaller effect of this shock on real activity.

C0615: **Rare disasters, the natural interest rate and monetary policy**
Presenter: **Alessandro Cantelmo**, Bank of Italy, Italy

The impact of rare disasters on the natural rate and macroeconomic conditions is evaluated by simulating a nonlinear New-Keynesian model. The model is calibrated using data on natural disasters in OECD countries. From an ex-ante perspective, disaster risk behaves as a negative demand shock and lowers the natural rate and inflation, even if disasters hit only the supply-side of the economy. These effects become larger and nonlinear if extreme natural disasters become more frequent, a scenario compatible with climate change projections. From an ex-post perspective, a disaster realization leads to a temporarily higher natural rate and inflation if supply-side effects prevail. If agents risk aversion increases temporarily, disasters may generate larger demand effects and lead to a lower natural rate and inflation. If supply-side effects dominate, the central bank could mitigate output losses at the cost of temporarily higher inflation in the short run. Conversely, under strict inflation targeting, inflation is stabilized at the cost of larger output losses.

C0494: **Systemic risk spillovers across the Euro Area**

Presenter: **Alexandros Skouralis**, Lancaster University; Central Bank of Ireland, United Kingdom

The high degree of financial contagion across the Euro area during the sovereign debt crisis highlighted the importance of systemic risk. We employ a GVAR model to analyse the systemic risk spillovers across the Euro area and to assess their role in the transmission of unconventional monetary policy. The results indicate a strong interconnectedness among core countries and also that peripheral economies have a disproportionate importance in spreading systemic risk. A systemic risk shock results in economic slowdown domestically and causes negative spillovers to the rest of the EMU economies. To examine how monetary policy impacts systemic risk, we incorporate high-frequency monetary surprises into the model. We find evidence of the risk-taking channel during normal times. In contrast, the relationship is being reversed in the period of the ZLB with expansionary shocks to result in a more stable financial system. Our findings indicate that the signalling channel is the main driver of this effect and that the initiation of the QE program boosts the economic activity but results in higher systemic risk. Finally, our results suggest that spillovers play an important role in the transmission of the monetary policy and that there is evidence of significant heterogeneity amongst countries' responses with core countries to benefit the most from changes in monetary policy.

Saturday 19.12.2020 13:35 - 15:15 Parallel Session D – CFE-CMStatistics

EO057 Room R11 STATISTICS FOR HILBERT SPACES I **Chair: Gil Gonzalez-Rodriguez**

E1058: **An ensemble method for multivariate functional data classification, with application to mouse movement trajectories**
Presenter: **Sonja Greven**, Humboldt University of Berlin, Germany
Co-authors: Amanda Fernandez-Fontelo, Felix Henninger, Pascal Kieslich, Frauke Kreuter

An ensemble method is presented for multivariate functional data classification that combines different semi-metric-based classifiers. We extend existing methods to the multivariate case and to further ensemble methods, and allow for scalar covariates. An R package implements the presented classification methods for multivariate functional data and trajectories in n dimensions. We apply our methods to the motivating application, to predict the difficulty of respondents while filling out a web survey using their computer mouse trajectories.

E0932: **A depth-based global envelope test with applications to biomedical functional data**
Presenter: **Sara Lopez Pintado**, Northeastern University, United States
Co-authors: Kun Qian

Functional data are commonly observed in many emerging biomedical fields and their analysis is an exciting developing area in statistics. The statistical analysis of functions can be significantly improved using non-parametric and robust estimators. New ideas of depth for functional data have been proposed in recent years and can be extended to image data. They provide a way of ordering curves or images from center-outward, and of defining robust order statistics in a functional context. We develop depth-based global envelope tests for comparing two-groups of functions or images. In addition to providing global p-values, the proposed envelope test can be displayed graphically and indicates the specific portion(s) of the functional data (e.g., in pixels or in time) that may have led to the rejection of the null hypothesis. We show in a simulation study the performance of the envelope test in terms of empirical power and size in different scenarios. The proposed depth-based global approach has good power even for small differences and is robust to outliers. The methodology introduced is applied to test whether children with normal and low birth weight have a similar growth pattern. We also analyzed a brain image data set consisting of positron emission tomography (PET) scans of severely depressed patients and healthy controls. The extension of the envelope test to multivariate functional data is explored.

E0767: **Characterizing smooth trends and irregular spikes in longitudinal data**
Presenter: **Huy Dang**, Penn State University, United States
Co-authors: Marzia Cremona, Francesca Chiaromonte

Many longitudinal data – in fields ranging from economics to the biomedical sciences, or the geosciences – comprise both smooth and irregular elements. We consider scenarios in which an underlying smooth curve is composed not just with Gaussian errors, but also with irregular spikes that (a) are themselves of interest, and (b) can negatively affect our ability to characterize the underlying curve. For such scenarios, we propose an approach that, combining regularized spline smoothing and an Expectation-Maximization algorithm, allows one to both identify spikes and estimate the smooth component. Imposing some assumptions on the error distribution, we prove the convergence of EM estimates to the true population parameters. Next, we demonstrate the performance of our proposal on finite samples and its robustness to assumptions violations through simulations. Finally, we apply our proposal to the analysis of two-time series data: one concerns the annual heatwaves index in the US over the past 100 years, the other concerns the weekly electricity consumption in Ireland. We characterize the underlying smooth trends in both dataset, as well as identify the irregular/extreme behaviors. In one of the applications, i.e. the annual heatwave data, the identified extremity is confirmed as a well-known extreme event.

E0919: **Sure Independence Screening (SIS) for multiple functional regression model**
Presenter: **Yuan Yuan**, Auburn University, United States
Co-authors: Nedret Billor

Due to rapid advancements in computer technology, high-dimensional, big and complex data, such as functional data where observations are considered as curves, have emerged from applications of biomedicine, chemometrics, engineering, and social sciences. Since functional data are inherently infinite-dimensional, variable selection problem in multiple functional regression model is, therefore, challenging and difficult. A novel approach is offered for functional feature selection under high-dimensional context based on Sure Independence Screening with functional predictors and scaler responses. High dimensionality means that $p = O(\exp(n^r))$, where p is the number of functional predictors, and n is the sample size. With simulation studies and real data application, the current method detects true feature set among thousands of functional features and show potential in high-dimensional functional classification as well.

EO738 Room R12 RECENT ADVANCES IN FUNCTIONAL TIME SERIES **Chair: Siegfried Hoermann**

E0351: **Detection of periodic signals in functional time series**
Presenter: **Vaidotas Characiejus**, University of California, Davis, United States
Co-authors: Siegfried Hoermann, Clement Cerovecki

A test is developed to detect periodic signals in functional time series when the length of the period is unknown. The observations are assumed to belong to an infinite-dimensional separable Hilbert space and the test is based on the asymptotic distribution of the maximum over all Fourier frequencies of the Hilbert-Schmidt norm of the periodogram operator. We show that under certain assumptions the appropriately standardized maximum of the periodogram belongs to the domain of attraction of the Gumbel distribution. The main ingredient of the proof is a recent Gaussian approximation. The results generalize a previous result to multivariate and functional time series. They also complement other results, where the length of the period is assumed to be known. We illustrate the usefulness of our test by examining the air quality data from Graz, Austria, and showing that our test is able to reveal a periodic component which is not a priori expected. We also demonstrate the finite sample performance of our test using a small simulation study.

E0477: **An autocovariance-based learning framework for high-dimensional functional time series**
Presenter: **Cheng Chen**, London School of Economics, United Kingdom
Co-authors: Jinyuan Chang, Xinghao Qiao

Many scientific and economic applications involve the analysis of high-dimensional functional time series, which stands at the intersection between functional time series and high-dimensional statistics gathering challenges of infinite-dimensionality with serial dependence and non-asymptotics. We model observed functional time series, which are subject to errors in the sense that each functional datum arises as to the sum of two uncorrelated components, one dynamic and one white noise. Motivated from a simple fact that the autocovariance function of observed functional time series automatically filters out the noise term, we propose an autocovariance-based three-step procedure by first performing autocovariance-based dimension reduction and then formulating a novel autocovariance-based block regularized minimum distance (RMD) estimation framework to produce block sparse estimates, from which we can finally recover functional sparse estimates. We investigate non-asymptotic properties of relevant estimated terms under such autocovariance-based dimension reduction framework. To provide theoretical guarantees for the second step, we also present a convergence analysis of the block RMD estimator. Finally, we illustrate the proposed autocovariance-based learning framework using applications of three sparse high-dimensional functional time series models. With derived theoretical results, we study convergence properties of the associated estimators.

E0626: **Factor modelling for functional time series**
Presenter: **Qingsong Wang**, Renmin University of China, China

Functional time series now arise in many scientific fields. We consider factor modelling for functional time series. To estimate both the number of factors and the factor loading space, we develop a fully functional method by performing an eigenanalysis for a nonnegative definite matrix, formed by cross- and auto-covariance functions and the double integration. We provide both an intuitive explanation from the regression perspective and theoretical support from the asymptotic perspective. The proposed method performs well, even when the number of functional variables is relatively large compared to the number of temporally dependent functional observations. Extensive simulation studies show that our method performs well in all cases. Finally, we demonstrate the superior sample performance of the proposed methods through a real data example.

E0666: **Factor-augmented smoothing model for raw functional data**
Presenter: **Yanrong Yang**, The Australian National University, Australia

The proposal is to model functional data as a mixture of a smooth function and a high dimensional factor component. The conventional approach to retrieving the smooth function from the raw data is through various smoothing techniques. However, the smoothing model is not adequate to recover the smooth curve or capture the data variation in some situations. These include cases where there is a large amount of measurement error, the smoothing basis functions are incorrectly identified, or the step jumps in the functional mean levels are neglected. To address these challenges, a factor-augmented smoothing model is proposed, and an iterative numerical estimation approach is implemented in practice. Asymptotic theorems are also established to demonstrate the effects of including factor struc- tures on the smoothing results. Specifically, we show that the smoothing coefficients projected on the complement space of the factor loading matrix are asymptotic normal. As a byproduct of independent interest, an estimator for the population covariance matrix of the raw data is presented based on the proposed model. Extensive simulation studies illustrate that these factor adjustments are essential in improving estimation accuracy and avoiding the curse of dimensionality. The superiority of our model is also shown using Canadian weather data and Australian temperature data.

EO508 Room R13 ADVANCES IN FINANCIAL AND PANDEMIC ECONOMETRICS Chair: Yi He

E0257: **Prediction analysis for extreme conditional quantiles through panel data quantile regression with individual effects**
Presenter: **Xuan Leng**, Xiamen University, China
Co-authors: Yanxi Hou

The aim is to study the estimation and prediction of extreme conditional quantiles of panel data based on individual-effect models. We propose a two-stage method, where the first stage is implemented based on the panel quantile models at an intermediate level and the second stage is based on the extrapolation method for an extreme level. The method relies on a set of second-order regular variation conditions of heteroscedastic extremes, which is used for the establishment of asymptotic normality for extreme conditional quantiles. Finally, a simulation study is evaluated to show the finite performance of the estimator, and a real data analysis are conducted for the illustration of the method.

E0259: **Weak factors robust Hansen-Jagannathan distance test**
Presenter: **Lingwei Kong**, University of Groningen, Netherlands

The Hansen-Jagannathan (HJ) statistic is one of the most dominant measures of model misspecification in asset pricing models. However, the conventional HJ specification test procedure has a poor finite sample performance, and it can be size distorted even in large samples when factors exhibit small correlations with asset returns. Applied researchers are likely to over-reject a model when it is correctly specified. We provide a novel model specification test, which is robust against the presence of weak factors and more powerful than the HJ test, and we also offer a novel robust risk premia estimator. The empirical application documents the non-reliability of the traditional HJ test since it may produce counter-intuitive results, when comparing nested models, by rejecting a four-factor model but not its embedded reduced three-factor model. At the same time, the proposed method is practically more appealing and show support for a four-factor model for Fama-French portfolios.

E0576: **Unified extreme value estimation for heterogeneous data**
Presenter: **Yi He**, University of Amsterdam, Netherlands
Co-authors: John Einmahl

The aim is to develop a universal econometric formulation of the empirical power laws possibly driven by parameter heterogeneity. The approach extends classical extreme value theory to specify the behavior of the empirical distribution of a general data set with possibly heterogeneous marginal distributions and a complex dependence structure. The main assumption is that in the intermediate tail, the empirical distribution approaches some heavy-tailed distribution with a positive extreme value index. In this setup the Hill estimator consistently estimates this extreme value index and, on a log-scale, extreme quantiles are consistently estimated. We discuss several model examples that satisfy our conditions and demonstrate in simulations how heterogeneity may generate the dynamics of empirical power laws. We observe a dynamic cross-sectional power law for the new confirmed COVID-19 cases and deaths per million people across countries. We show that this international inequality is largely driven by the heterogeneity of the countries scale parameters.

E0805: **Alternating pruned dynamic programming for multiple epidemic change-point estimation**
Presenter: **Zifeng Zhao**, University of Notre Dame, United States

The problem of multiple change-point detection for a univariate sequence under the epidemic setting is studied, where the behavior of the sequence alternates between a common normal state and different epidemic states. This is a non-trivial generalization of the classical (single) epidemic change-point testing problem. To explicitly incorporate the alternating structure of the problem, we propose a novel model selection based approach for simultaneous inference on both change-points and alternating states. Using the same spirit as profile likelihood, we develop a two-stage alternating pruned dynamic programming algorithm, which conducts efficient and exact optimization of the model selection criteria and has $O(n^2)$ as the worst-case computational cost. As demonstrated by extensive numerical experiments, compared to classical general-purpose multiple change-point detection procedures, the proposed method improves accuracy for both change-point estimation and model parameter estimation. We further show promising applications of the proposed algorithm to multiple testing with locally clustered signals, and demonstrate its advantages over existing methods in large scale multiple testing, in DNA copy number variation detection, and in an oceanographic study.

EO131 Room R14 NEW PROPOSALS FOR THE ANALYSIS OF ORDINAL AND MIXED-TYPE DATA Chair: Cristina Mollica

E0791: **The optimal number of response alternatives for a rating scale**
Presenter: **Maria Iannario**, University of Naples Federico II, Italy
Co-authors: Anna Clara Monti, Pietro Scalera

The most widespread models for the analysis of ordinal responses are the cumulative models with the proportional assumption, which implies that the covariates have the same effect on the cumulative odds regardless of the category of the response. This assumption allows for generating ordinal responses with a different number of categories from the same latent variable. By exploiting this latter property, the analysis investigates the asymptotic efficiency of the estimators and the power of the tests when the number of response alternatives varies. The impact of collapsing categories is also considered with special focus on dichotomization. A discussion in case of Additive Location-Shift Models is dealt with in its initial stage.

E0744: **Computer skills for e-learning activities of university students: A posetic approach**
Presenter: **Alberto Arcagni**, Sapienza University of Roma, Italy
Co-authors: Cristina Mollica, Marco Fattore

The motivation comes from an international research project promoted by the University of Ljubljana (Slovenia), that conducted a large-scale survey on the COVID-19 effects on higher education students. The survey contains different ordinal variables and other categorical variables describing the academic environment. This data structure configures a partially ordered set (poset); therefore, the Fuzzy First Order Dominance (FFOD) method is particularly indicated to analyze it. By focussing on the Italian academic context, the FFOD is employed to explore the responses provided by a sample of students enrolled in degree or doctorate courses of Italian universities. During the lockdown, these students were asked to score their computer skills related to the e-learning activities. The considered ordered variables concern students' expertise in both managing digital communication/information channels and using more specific platforms for online lectures and academic support to their studies. The ratings are analyzed with the FFOD method to accomplish a comparative evaluation of different study courses and assess possible relationships of the resulting orderings with other academic variables typically impacting on university performance.

E0894: **An IRT model for evaluating university students' satisfaction about the on-line activities during the COVID-19 pandemic**
Presenter: **Serena Arima**, University of Salento, Italy
Co-authors: Cristina Mollica

The transition from the second to the third decade of the new century was sanctioned by the serious health crisis caused by COrona VIrus Disease (COVID-19). In order to contain the spiral of infections and thus avoid the collapse of the health structures, governments imposed a series of restrictive measures which, during the 2020 springtime, culminated in a period of lockdown. As a consequence, the traditional academic institutions reformulated their training activities and administrative services remotely, through the massive use of digital communication technologies. Motivated by an international project promoted by the University of Ljubljana in Slovenia, aimed at studying the COVID-19 effects on the life of higher education students, the performance evaluation of the Italian academic education institutions in coping the COVID-19 pandemic is considered, based on the satisfaction experienced during the lockdown by a sample of university students enrolled in a degree or doctorate courses. We propose a Bayesian item response model, a mixed-effects graded response model, accounting for the different nature of the possible answers. The latent trait, e.g. the satisfaction of each student, is hierarchically modelled as a function of specific characteristics of the corresponding university. Universities are then ranked using stochastic dominance criteria.

E0957: **A simple generative model for rank ordered data with ties**
Presenter: **Daniel Henderson**, Newcastle University, United Kingdom

A model for ranked ordered data which naturally accommodates ties is proposed. The model is inspired by the exponential latent variable formulation of the Plackett-Luce model and can be seen as its discrete counterpart. Specifically, the generative model assumes that the data arise from independent geometric latent variables. The latent variables can be integrated out of the model analytically, resulting in a simple likelihood function that facilitates straightforward inference. With a focus on Bayesian inference, a simple Gibbs sampling algorithm is presented. Several extensions of the basic model are considered.

EO285 Room R15 DATA SCIENCE METHODS FOR INTELLIGENT TEXT AND VIDEO PROCESSING Chair: Roy Welsch

E1106: **Intelligent asset management and NLP**
Presenter: **Frank Xing**, Nanyang Technological University, Singapore

Asset allocation models consider financial variables that are traditionally computed from numerical format data. Whereas the amount of unstructured text data has surged in the past decades, there is no place to accommodate such data in asset allocation models. We propose a possible solution to integrate such data, thus make asset management more intelligent. The solution links the text data to the old variables using natural language processing techniques. We elaborate on two examples: one is to approach asset return distribution via Bayesian revision using asset-specific news sentiment, the other is to model dependence using asset correlations estimated from the semantic information of company descriptions. In various simulations and experiments, introducing text data appears to be beneficial. Specifically, market sentiment views increase major performance indicators, such as CAGR and Sharpe ratio, by 5% to 20%; the semantic vine (used for covariance matrix estimation) also has a 98% chance to outperform an arbitrary vine for portfolio optimization. Moreover, the results could potentially improve as more sophisticated NLP methods are adapted. Similar opportunities might exist for other financial variables needed for asset management.

E1165: **Applications of machine learning in data and text**
Presenter: **Hongyan Cui**, Beijing University of Posts and Telecommunications, China
Co-authors: Ensen Wu, Zhenwei Sun, Gangkun Wang, Songdeng Hui

In recent decades, machine learning has played a significant role in a wide spectrum of industries due to its excellent processing ability. According to different types of data and tasks, machine learning applications can be divided into different domains, such as structure data classification, NL2SQL tasks and Semi-supervised short text classification etc. However, there are still many problems in these domains that need to be further explored, such as imbalanced classification problem, multi-task learning problems and sparsity and limited labeled data problems. Examples of our recent researches in machine learning applications in the above domains are described. Firstly, we will introduce the application of Generative Adversarial Networks and Autoencoders to solve imbalanced classification problem in structure data classification. Secondly, we will describe a pre-trained language model structure that performs well in NL2SQL tasks. Finally, we will introduce a graph neural network and self-training method, which improves the classification performance of short text.

E1173: **Instability of deep learning models in industrial NLP problems**
Presenter: **Helen Xie**, Apple, United States

While many industrial companies turn to deep learning and/or machine learning models to achieve smart solutions for complicated NLP problems, a stable model is desired but sometimes hard to reach. Such a problem might never show up in academia, but only industrial. Several main reasons can be the design of infrastructure, algorithm and GPU related issues, and inconsistency in human-labelled data. For industrial, infrastructure is usually tuned to be able to hold trillions of traffics smoothly and efficiently but not highly consistently. Implementation of the experimental level algorithm to production has to be re-shaped, and stability usually is not the focus. Updating to the newest GPU can be slow. Human labels can constantly be changing as the algorithm changes. As a result, several statistic methods are applied to restrict models with high fluctuations (CI 95%) to be shipped. Unfortunately, even with such restrictions, instability within models still exists, but from 32.2% to 23.4% in accuracy, using simulated models and data.

E1175: **Predicting video engagement using heterogeneous DeepWalk**
Presenter: **Iti Chaturvedi**, James Cook University, Australia

Video engagement is important in online advertisements where there is no physical interaction with the consumer. This can help identify advertisement frauds where a robot runs fake videos behind the name of well-known brands. Engagement can be directly measured as the number of seconds after which a consumer skips an advertisement. Furthermore, we leverage the fact that videos shown on the same channel have similar viewing behavior. Hence, we use a graph-embedding model called DeepWalk to determine the clusters of videos with high engagement in a particular channel. The learned embedding is able to identify viewing patterns of fraud and popular videos. In order to assess the impact of a video we

also consider how the view counts increase or decrease over time. This results in a heterogeneous graph where an edge indicates similar video engagement or history of view counts between two videos. Since it is difficult to find labeled samples for 'fraud' video, we leverage a one-class model that can determine 'fraud' videos with an outlier or abnormal behaviour. The proposed model outperforms baselines in regression error by over 20%.

EO532 Room R16 NETWORK DIFFUSION MODELLING AND STOCHASTIC OPTIMIZATION Chair: Sophie Dabo

E0216: **Financial network discovery from restricted vector autoregression: New model and application**
Presenter: **Stefano Nasini**, IESEG School of Management, France
Co-authors: Massimiliano Caporin, Deniz Erdemlioglu

A new VAR model is developed that uncovers the interconnectedness among financial assets by aggregating realized trading information from intraday data. With a limited number of parameters, the presented model can accommodate the dynamic interactions in large panels of financial assets and realised measures (such as returns, volatilities, etc.). We propose an alternating direction method for maximum likelihood estimation of this type of VAR family. Using daily cross-sectional data on 1095 individual firms over fifteen years, we show that the full model estimation is reliable and computationally feasible. Our empirical results support the practical effectiveness of our model for ranking systemically important financial institutions.

E0487: **Endogenous technology sharing in decentralized production**
Presenter: **Marijn Verschelde**, IESEG School of Management, France
Co-authors: Stefano Nasini, Bruno Merlevede

The joint nature of knowledge-based assets drives firms to pursue multi-plant production. By facilitating knowledge sharing across plants, integrated firms have higher returns on investment for intangibles in comparison to single-plant firms. We study the endogenous choice of intangible assets by the parent firm, organizing decentralized multi-plant production. A distinguishing feature of our framework is that we allow for anticipation by the parent of the affiliates' best response to the technology transfer and transfer price, set by the parent. Our approach is general in the sense that we cover both horizontal and vertical production and allow for dynamic optimization by the follower. We implement the leader-follower structure by the use of a bi-level optimization framework wherein the parent acts as a newsvendor. For horizontally organized firm structures, we recover the optimal solution, and for vertically organized firm structures, we recover tight lower and upper bounds. We show the empirical applicability by the use of a customized version of the Orbis dataset that focuses on European firms with complete parent-affiliate balance sheet information at the firm-year level. Our advocated approach nicely recovers stylized facts and pinpoints profit gains that originate from anticipating the best response of the follower.

E0545: **A new class of decentralized portfolio optimization problems for risk diffusion**
Presenter: **Nessah Rabia**, IESEG School of Management, France
Co-authors: Stefano Nasini, Francisco Benita

In the context of risk management, a decentralized investment problem for risk minimization in multiple markets is modeled, where a single investor delegates the portfolio decisions to risk-minimizers intermediaries. The general optimization problem consists of a single-leader multi-follower game, whose numerical solution is addressed by taking advantage of uncovered properties of the risk diffusion measurements. We show that under mild assumptions, this problem can be reformulated as a non-linear knapsack problem. We numerically assess the property of the solution using large-scale stock returns data from U.S. listed enterprises.

E0880: **An asymptotic approximation for the extended Bass diffusion model**
Presenter: **Sophie Dabo**, University of Lille, France
Co-authors: Ringo Thomas Tchouya, Stefano Nasini

The continuous-time dynamic problem of diffusion in bipartite networks is studied, and a model is proposed in which an influencer node is responsible for the propagation of a binary state towards multiple independent populations. We deduce an asymptotically correct approximation scheme for the underlying differential equation, that allows embedding the model into computationally treatable estimation approach. This supports the estimation of the unknown parameters of this model, using observed time series data. A comprehensive numerical study is presented to assess the validity of the proposed approach.

EO041 Room R17 ADVANCES IN HIGH DIMENSIONAL TIME SERIES MODELS Chair: George Michailidis

E0332: **Simultaneous prediction intervals for high-dimensional vector autoregression**
Presenter: **Sayar Karmakar**, University of Florida, United States
Co-authors: Mengyu Xu

Simultaneous prediction intervals for high-dimensional vector autoregressive processes are studied. Motivated from an online change-point problem, we wish to construct prediction intervals for one-step-ahead predictions in a high-dimensional VAR process. A de-biased calibration is used to post-regularize the lasso estimation, and a novel Gaussian-multiplier bootstrap-based method is developed for one-step-ahead prediction. The asymptotic coverage consistency of the prediction interval is obtained. We also substantiate the theoretical result by some simulations for evaluating finite sample performance and show some real data analysis. Our simulated results show considerably good performance even in situations where p is much larger than n where most of the previous literature only focused on situations where p grows but remains less than n.

E0514: **Sparse locally-stationary wavelet processes**
Presenter: **Alex Gibberd**, Lancaster University, United Kingdom

Traditional estimation of graphical model structure assumes that data is drawn i.i.d. from some underlying homogeneous population. Across the multiple domains of economics, neuroscience, and genetics, these assumptions have been recognised as unrealistic and given rise to model extensions that permit heterogeneous populations. We will discuss a further extension of graphical dependency modelling that allows us to model both temporal and cross-sectional dependency structure. Specifically, we will introduce a new class of sparse locally-stationary wavelet (sLSW) processes and suggest efficient regularised M-estimators for their identification. Constructed in relation to families of non-decimated wavelets, these processes can represent a wide variety of time-series with cross-sectional (and time-varying) dependence. One important and useful consequence of the wavelet modelling approach is that dependency structure can be isolated to distinct scale (or aggregation) levels.

E0904: **Efficient hyperparameter selection for penalised regression using communicating gradient descent algorithm**
Presenter: **Sandipan Roy**, University of Bath, United Kingdom
Co-authors: Stephane Chretien, Alex Gibberd

In high-dimensional regression problems, statistical learning often relies on sparsity assumptions on the regression vector, which are often enforced using L1-type penalties in the estimation stage. One of the main difficulties in solving such penalised problems is to calibrate the so-called relaxation parameter or hyperparameter accurately. Many different methods are available for the hyperparameter selection problem: Bayesian Optimisation, Cross-Validation, Multi-Armed Band algorithms, etc. We propose a new methodology based on running communicating pools of incremental gradient algorithms in parallel, each of them corresponding to a specific value of the hyperparameter on a grid. Each time a new data is observed, the prediction performance of the different values of the hyperparameter can be compared using a Follow the Leader scheme. Theoretical guarantees

are provided for our method, showing the power of communication for this problem. The results are illustrated with numerical experiments showing that our simple selection rule can achieve prediction results comparable to state of the art, at a lower computational cost

E0946: **Anomaly detection for high-dimensional linear regression models with possible temporal dependence**
Presenter: **Abolfazl Safikhani**, University of Florida, United States
Co-authors: Yue Bai

Detecting structural breaks in high-dimensional linear regression models is a challenging task due to the existence of an unknown number of such breaks, unknown location of breaks, and unknown high-dimensional model parameters. A general methodology for handling this problem will be presented which can cover a wide range of statistical models including mean shift models, Vector Auto-Regressive Models (VARs), and Gaussian graphical models. The proposed algorithm consists of (1) applying blocked fused lasso to identify potential locations of breaks; (2) evaluate the magnitude of changes and apply thresholding to keep only the large enough jumps; (3) apply k-means clustering to the location of large jumps to select clusters around true breaks; (4) exhaustive search within each selected cluster to locate the breaks. Consistency for estimating the number of breaks, their locations and model parameters is verified under mild conditions satisficed in many well-known high-dimensional statistical models. The proposed method performs well in synthetic and real data applications while outperforming some competing methods in the literature.

EO542 Room R18 NEW ADVANCES IN BIOMEDICAL DATA ANALYSIS Chair: Yichuan Zhao

E0341: **Accelerating health research with electronic health records data**
Presenter: **Rebecca Hubbard**, University of Pennsylvania, United States

Using data generated as a by-product of electronic interactions, including electronic health records (EHR), social media data, and data from wearable devices, has the potential to accelerate research on health and healthcare vastly. Statistical insights on sampling and inference are key to drawing valid conclusions based on these messy and incomplete data sources. We will use previous research on EHR-based phenotyping to motivate a discussion of the roles of informatics, statistics, and data science in the process of learning from EHR data. EHR-based phenotyping is hampered by complex missing data patterns and heterogeneity across patients and healthcare systems, features which have been largely ignored by existing phenotyping methods. As a result, not only are EHR-derived phenotypes imperfect, but they often feature exposure-dependent differential misclassification, which can bias results towards or away from the null. We will discuss novel and existing approaches to EHR-based phenotyping, as well as statistical methods to correct for phenotyping error in analyses. The overall goal is to use the example of phenotyping to illustrate the unique contribution of statistics to the process of generating evidence from modern data sources.

E0291: **Semiparametric trend analysis for stratified recurrent gap times under weak comparability constraint**
Presenter: **Peng Liu**, University of Kent, United Kingdom

Recurrent event data are frequently found in many clinical trial studies and medical research, where each subject encounters more than one sequential event. A much-discussed aspect of the recurrent events is the presence or absence of the time trend, where trend refers to a systematic variation among the length of the sequential gap times, which can be used as a measure of disease progression. Under the accelerated failure time (AFT) model, a comparability concept has been previously proposed to estimate the trend among sequential gap times (slope parameter). Each individual gap time has the same distribution subject to the comparability constraint, and thus the estimation can be easily conducted. However, their comparability is a strong constraint. We propose a weak comparability constraint under the previos assumption for the AFT model. The estimator will utilize more data due to weaker constraint, and thus it will result in a more efficient estimate for the slope parameter in the AFT model. Monte Carlo simulation is performed to validate the effectiveness of the new method. The method is applied to the HIV Prevention Trial Network (HPTN) 052 study.

E0851: **Sensitivity analysis: E-value based on direct adjusted survival probabilities**
Presenter: **Zhang Mei-Jie**, Medical College of Wisconsin, United States
Co-authors: Cheng Zheng, Zhenhuan Hu

The estimated treatment effect from an observational study may be biased, and potential unmeasured confounding bias could be a central limitation of observational studies. Sensitivity analysis is useful in assessing the robustness of the estimated treatment effect based on a statistical analysis of observational data. Recently, a new sensitivity analysis method called E-value has been proposed, which measures potential inference of unmeasured confounding in observational studies. E-value is defined as the minimum strength of association of an unmeasured confounder would need to explain away a treatment-outcome association. A common approach to testing for differences in survival rates between two therapies while adjusting for prognostic factors is to compare the direct adjusted survival curves. The direct adjusted survival probabilities are estimated by taking the average of the individual predicted survival curves over the entire study cohort, which control for the possible imbalance of patient characteristics between treatment groups. We suggest a sensitivity analysis using an E-value for comparing two direct adjusted survival probabilities. A real stem cell transplant data example illustrates the practical utility of the proposed method.

E0361: **Reduce the computation in jackknife empirical likelihood for comparing two correlated Gini indices**
Presenter: **Yichuan Zhao**, Georgia State University, United States
Co-authors: Kangni Alemdjrodo

The Gini index has been widely used as a measure of income (or wealth) inequality in social sciences. To construct a confidence interval for the difference of two Gini indices from the paired samples, a profile jackknife empirical likelihood after maximization over a nuisance parameter has been used, and a Wilks' theorem has been established. However, profiling could be very expensive. We propose an alternative approach of the jackknife empirical likelihood method to reduce the computational cost. We also investigate the adjusted jackknife empirical likelihood and the bootstrap-calibrated jackknife empirical likelihood to improve coverage accuracy for small samples. Simulations show that the proposed methods perform better than the previous methods in terms of coverage accuracy and computational time. Two real data applications proved that the proposed methods work perfectly in practice.

EO766 Room R19 RECENT ADVANCES IN STATISTICAL METHODS FOR MOBILE HEALTH Chair: Walter Dempsey

E0725: **Combining model-based and model-free methods for estimating interventions with indefinite horizons**
Presenter: **Eric Laber**, North Carolina State University, United States
Co-authors: Owen Leete

Mobile-health (mHealth) holds tremendous potential as a means of delivering precision medicine at scale. mHealth-based precision medicine strategies seek to tailor intervention decisions to the unique health trajectory of each patient. An optimal intervention strategy maximizes some cumulative measure of patient health over the intervention period. Most commonly used methods for estimation of an optimal intervention strategy can be broadly characterized as either model-based, in which the underlying generative model is estimated and subsequently used to identify an optimal strategy using simulation (g-computation), or model-free, in which semi-parametric estimating equations are used to identify an optimal regime. Model-based methods impose a structure that reduces variance but is subject to misspecification, which may induce bias. In contrast, model-free methods impose less structure which reduces the risk of bias but may increase variance. We propose a method that combines model-free and model-based estimators which is consistent if one (but not necessarily both) is correctly specified and efficient if both are correctly specified. Empirical results show the proposed approach mitigates the risk of misspecification and yields better patient outcomes than competing methods.

E1066: **Modelling wearable data via quantile-based distributional data analysis**
Presenter: **Vadim Zipunnikov**, Johns Hopkins University, Bloomberg School of Public Health, United States

With the advent of continuous health monitoring via wearable devices and digital sensors, users now generate their own unique stream of continuous data such as minute-by-minute heart rate or blood pressure. Aggregating these streams into scalar summaries ignores the distributional nature of data and often leads to the loss of critical information. We propose to capture the distributional nature of wearable data via user-specific quantile functions and develop flexible methods to analyze these functions as hybrid functional-distributional data. Traditional approaches of using a single distributional summary such as moments or extremes become special cases of the proposed method. Specifically, the use of L-moments to represent quantile-functions via interpretable decompositions allows us to define an interpretable distance between distributional observations. The quantile functions and L-moments are shown to be flexible to be employed within generalized scalar-on-function regression models and for analyzing joint and individual sources of variation of multimodal data. The proposed methods are illustrated in a study of the association between accelerometry-derived digital gait biomarkers with Alzheimer's disease (AD) and cognitive functioning. Our methods allow digital biomarkers of gait such as step velocity, cadence, stride regularity and mean stride time to be highly discriminatory for subjects with AD and impaired cognitive performance.

E0812: **Latent variable regression analysis of longitudinal multivariate data with irregular and informative observation times**
Presenter: **Zhenke Wu**, University of Michigan, United States

In many mobile health studies, phone surveys such as Ecological Momentary Assessments (EMA) are increasingly adopted because they are less susceptible to recall bias and are sensitive to contextual factors. For example, they hold great potentials in smoking cessation studies to probe subjects' time-varying psychological states such as vulnerability (risk for lapse) and receptivity (ability and willingness to engage with self-regulatory activities). Inference and prediction of these states may inform just-in-time adaptive intervention development. However, the observation times of these EMAs may correlate with survey responses. For instance, some EMAs are delivered and answered with lower positive emotions when they were triggered by a recent smoking episode detected by on-body sensors. Such dependence must be accounted for to obtain valid inference. We propose a latent variable regression approach for longitudinal multivariate discrete data analysis with irregular and informative observation times. The goal is to infer the distribution of scientifically meaningful latent variables over time as a function of covariates. The observed dependence between the survey responses and observation times is assumed to be induced by unobserved random effects and observed covariates. We demonstrate the utility of the proposed model through simulation studies and an analysis of data from Break Free study among African Americans who attempt to quit smoking.

E0713: **Personalized policy learning using longitudinal mobile health data**
Presenter: **Ken Cheung**, Columbia University, United States
Co-authors: Xinyu Hu, Min Qian, Bin Cheng

The personalized policy learning problem is addressed by using longitudinal mobile health application usage data. Personalized policy represents a paradigm shift from developing a single policy that may prescribe personalized decisions by tailoring. Specifically, we aim to develop the best policy, one per user, based on estimating random effects under a generalized linear mixed model. With many random effects, we consider new estimation method and penalized objective to circumvent high-dimension integrals for marginal likelihood approximation. We establish consistency and optimality of our method with endogenous app usage. We apply the method to develop personalized push (prompt) schedules in 294 app users, intending to maximize the prompt response rate given past app usage and other contextual factors. We found the best push schedule given the same covariates varied among the users, thus calling for personalized policies. Using the estimated personalized policies would have achieved a mean prompt response rate of 23% in these users at 16 weeks or later: this is a remarkable improvement on the observed rate (11%), while the literature suggests 3%-15% user engagement at 3 months after download. The proposed method compares favorably to existing estimation methods, including using the R function glmer in a simulation study.

EO552 Room R20 ANALYZING COMPOSITIONAL, DISTRIBUTIONAL AND RELATIVE ABUNDANCE DATA II Chair: Karel Hron

E0988: **Seasonal variability in stream water chemistry assessed through a principal balances based approach**
Presenter: **Caterina Gozzi**, University of Florence, Italy
Co-authors: Antonella Buccianti

In modern times, climate change is significantly impacting freshwater resource availability. A warmer climate accelerates the water cycle, changing hydrological pathways, increasing evaporation, altering rainfall intensity, distribution and runoff regime. The implications on seasonal variability in stream water chemistry are significant, and the use of new compositional methods is mandatory to capture the relative behavior of solutes from a holistic perspective. A Principal Balances based approach was used to investigate the seasonal variability of 222 river waters sampled along the entire Tiber River catchment (central Italy). The same sequence of isometric log-ratio coordinates obtained for the whole dataset was applied separately to data from different flow regimes, evaluating changes due to seasonality. Principal Balances were compared utilizing a joint analysis of the kernel density distributions, and relevant deviations were highlighted using bar-plots. Results indicate a higher variability in drought periods and an influence on water chemistry of silicate weathering processes, which are strongly dependent on runoff fluctuations. The method proved to be effective and provides new insights for seasonal analysis on compositional data.

E0985: **Weighting of parts in compositional data and its applications**
Presenter: **Karel Hron**, Palacky University, Czech Republic
Co-authors: Alessandra Menafoglio, Javier Palarea-Albaladejo, Peter Filzmoser, Renata Talska, Juan Jose Egozcue

It often occurs in practice that it is sensible to give different weights to the variables involved in multivariate data analysis. The same holds for compositional data as multivariate observations carrying relative information, such as proportions or percentages. It can be convenient to apply weights to, for example, better accommodate differences in the quality of the measurements, the occurrence of zeros and missing values, or generally to highlight some specific features of compositional variables (i.e. parts of a whole). The characterisation of compositional data as elements of a Bayes space enables the definition of a formal framework to implement weighting schemes for the parts of a composition. This is formally achieved by considering a reference measure in the Bayes space alternative to the common uniform measure via the well-known chain rule. Unweighted centred log-ratio (clr) coefficients and isometric log-ratio (ilr) coordinates then allow representing compositions in the real space equipped with the (unweighted) Euclidean geometry, where ordinary multivariate statistical methods can be used and interpreted. We present these formal developments and use them to introduce a general approach to weighting parts in compositional data analysis. We demonstrate its practical usefulness on simulated and real-world data sets in the context of the earth sciences.

E0578: **Approximation of density functions using compositional splines**
Presenter: **Jitka Machalova**, Palacky University, Czech Republic
Co-authors: Karel Hron, Renata Talska

Probability density functions result in practice frequently from the aggregation of massive data, and their further statistical processing is thus of increasing importance. However, 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 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. Aim of the contribution is to introduce a new class of B-splines within the Bayes space

methodology which is suitable for representation of density functions. Accordingly, the original densities are expressed as real functions using the centred log-ratio transformation, and optimal smoothing splines with B-spline basis honoring the resulting zero-integral constraint are developed.

E1000: **Compositional scalar-on-function regression with a geological application**

Presenter: **Ivana Pavlu**, Palacky University Olomouc, Czech Republic

Co-authors: Renata Talska, Karel Hron, Daniel Simicek, Ondrej Babek

Regression between a real response and a density function as covariate has many practical motivations, e.g., to find a relationship between the geochemical composition of sediments and the distribution of particle sizes in soil (particle size distribution, PSD). In this case, the explanatory variable can be described in the form of the probability density function. At the same time, the response is a real variable (a meaningful scale-free representation of the original concentrations using log-ratios). Due to the relative character of densities, the Bayes space methodology was employed. Specifically, the centred log-ratio (clr) transformation played the role to represent the PSDs (densities) in the standard L^2 space. The idea of smoothing splines was used to represent the discretized input densities while fulfilling the zero-integral constraint imposed by the clr transformation. The resulting regression parameters (densities) can be interpreted in both the original and clr space; however, in the latter, the interpretation is more straightforward. The newly developed regression model, called compositional scalar-on-function regression, was examined with both simulated observations and real-world geological data; the latter were collected at four sites in the Czech Republic (Brodek u Prerova, Dobsice, Ivan, Rozvadovice). The regression model has proven to be a good tool for linking the grain size effect with geochemical signals (provenance, weathering, diagenesis, etc.).

EO105 Room R21 RECENT ADVANCE IS NETWORK ANALYSIS Chair: Jiashun Jin

E0264: **Statistics for statisticians**

Presenter: **Jiashun Jin**, Carnegie Mellon University, United States

A data set for the publications of statisticians has been collected and cleaned. It consists of titles, authors, author affiliations, abstracts, MSC numbers, keywords, reference, and citation. It counts of 83,661 papers published in 36 journals in statistics, probability, and related field, spanning 41 years. The data set motivates an array of interesting problems. We will discuss paper counts, most cited authors and papers, journal ranking, text mining, network analysis, and citation prediction. For text mining, we use the paper abstracts in our data set as the text documents, and focus on how to use the estimated topic weights to study the research patterns of individual authors. For network analysis, we focus on hierarchical community detection, membership estimation, and especially how to characterize the research trajectories of a subset of selected statisticians over the years.

E0265: **A front knowledge mapping analysis of international statistical research based on DCMM**

Presenter: **Xing Wang**, Renmin University of China, China

Based on 31,681 academic papers from 22 authoritative journals of Statistics from 2008 to 2018 collected by Web of Science, a keyword co-occurrence network is built to explore the frontier knowledge map. The keywords structure and interaction mechanism of international Statistics are explored to reveal the connection mechanism of micro-key concepts in different fields within the discipline. By using the Degree-Corrected Mixed-Membership (DCMM) model and its mix-SCORE algorithm, the probability of keywords belonging to the different frontier subjects is estimated. Keywords network topology structure of keyword co-occurrence network is studied. With the bibliometrics experience, the fundamental law of knowledge development of Statistics is revealed, including research fields of Statistics, research hotspots of famous international universities, and the hotspots' growth path. The results illustrate that the influence of the keyword co-occurrence network structure is more significant than that of the keyword itself. The connection mechanism and inheritance mechanism of different universities have a substantial impact. The keyword co-occurrence networks of well-known international universities have various preferential connection mechanisms and diversified academic ecology. The keyword co-occurrence tends to be connected with newer hotspots nodes.

E0339: **Testing community structure for hypergraphs**

Presenter: **Yang Feng**, NYU, United States

Many complex networks in the real world can be formulated as hypergraphs, where community detection has been widely used. However, the fundamental question of whether communities exist or not in an observed hypergraph remains unclear. The aim is to tackle this important problem. Specifically, we systematically study when a hypergraph with community structure can be successfully distinguished from its Erdos-Renyi counterpart, and propose concrete test statistics when the models are distinguishable. The main contribution is threefold. First, we discover a phase transition in the hyperedge probability for distinguishability. Second, in the bounded-degree regime, we derive a sharp signal-to-noise ratio (SNR) threshold for distinguishability in the special two-community 3-uniform hypergraphs and derive nearly tight SNR thresholds in the general two-community m-uniform hypergraphs. Third, in the dense-degree regime, we propose a computationally feasible test based on sub-hypergraph counts and obtain its asymptotic distribution and analyze its power. The results are further extended to non-uniform hypergraphs in which a new test involving both edge and hyperedge information is proposed. The proofs rely on Jansons contiguity theory, a high-moments-driven asymptotic normality result, and a truncation technique for analyzing the likelihood ratio.

E0760: **Estimating the number of communities by stepwise goodness-of-fit**

Presenter: **Tracy Ke**, Harvard University, United States

Co-authors: Jiashun Jin, Tracy Ke, Shengming Luo, Minzhe Wang

Given a symmetric network with n nodes, how to estimate the number of communities K is a fundamental problem. We propose Stepwise Goodness-of-Fit (StGoF) as a new approach to estimate K. For $m = 1, 2, \ldots$, StGoF alternately uses a community detection step and a goodness-of-fit step. We use SCORE for community detection, and propose a goodness-of-fit (GoF) measure. We show that the GoF statistic converges to $N(0,1)$ when $m < K$ and diverges to infinity in probability when $m = K$. Therefore, with a proper threshold, StGoF terminates at $m = K$ as desired. We consider a broad setting where we allow severe degree heterogeneity, a wide range of sparsity, and especially weak signals. In particular, we propose a measure for signal-to-noise ratio (SNR) and show that there is a phase transition: when SNR $\to 0$ as $n \to \infty$, consistent estimates for K do not exist, and when SNR tend to infinity, StGoF is consistent, uniformly for a broad class of settings. In this sense, StGoF achieves the optimal phase transition. Stepwise testing algorithms of a similar kind are known to face analytical challenges. We overcome the challenges by using a different design in the stepwise algorithm and by deriving sharp results in the under-fitting case ($m < K$) and the null case ($m = K$). The key to our analysis is to show that SCORE has the Non-Splitting Property (NSP). The NSPis non-obvious, so additional to rigorous proofs, we also provide an intuitive explanation.

E1186: **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. This computational challenge has been previously addressed with a fast pseudo-likelihood approach for fitting stochastic block models to large sparse networks. However, this approach does not have 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 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.

EO684 Room R22 BAYESIAN INVERSE PROBLEMS Chair: Natalia Bochkina

E0201: **Seismic imaging and uncertainty assessment using variational methods**
Presenter: **Xin Zhang**, University of Edinburgh, United Kingdom
Co-authors: Andrew Curtis

In a variety of geoscientic applications, maps of subsurface properties together with the corresponding maps of uncertainties to assess their reliability are required. Seismic tomography is a method that is widely used to generate those maps. Since tomography is significantly nonlinear, Monte Carlo sampling methods are often used for this purpose, but they are generally computationally intractable for large data sets and high-dimensionality parameter spaces. Variational methods solve the Bayesian inference problem as an optimization problem, yet still provide fully probabilistic results. We test two variational methods: automatic differential variational inference (ADVI) and Stein variational gradient descent (SVGD). We use them to solve both travel time tomography and full waveform inversion problems. The results show that variational inference methods can produce accurate approximations to the results of Monte Carlo sampling methods at significantly lower computational cost, provided that gradients of parameters with respect to data can be calculated efficiently. We therefore contend that variational methods may have greater potential to extend probabilistic analysis to higher dimensional tomographic systems than current Monte Carlo methods.

E0489: **Bayesian estimation and comparison of conditional moment models**
Presenter: **Anna Simoni**, CNRS - CREST, France
Co-authors: Siddhartha Chib, Minchul Shin

The focus is on the Bayesian analysis of 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 (ETEL) 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). The posterior distribution is shown to satisfy the Bernstein-von Mises theorem, subject to a growth rate condition on the number of approximating functions, even under misspecification of the conditional moments. A large-sample theory for comparing different conditional moment models is developed. The central result is that the marginal likelihood criterion selects the model that is less misspecified. Examples are provided to illustrate the framework and results.

E0583: **Rates of posterior contraction for spatially inhomogeneous unknowns**
Presenter: **Sergios Agapiou**, University of Cyprus, Cyprus
Co-authors: Masoumeh Dashti, Tapio Helin

Some recent results on frequentist convergence rates of the posterior distribution in nonparametric settings will be discussed. In particular, we will consider a class of prior distributions, with tails between Gaussian and exponential, called p-exponential priors. This class includes the 1-Besov priors, which are especially popular in the applied Bayesian inverse problems community due to their sparsity promoting and edge-preserving properties. We will present a general theory, and then we will focus on the white noise model with alpha-regular p-exponential priors. We will discuss contraction rates over Besov regularity of the truth which suggest that when interested in spatially inhomogeneous unknown functions, in terms of posterior contraction, it is preferable to use Laplace rather than Gaussian priors.

E0676: **Adaptation in Bayesian inverse problems using an empirical Bayes approach**
Presenter: **Natalia Bochkina**, University of Edinburgh, United Kingdom
Co-authors: Jenovah Rodrigues

A sequence space formulation of Bayesian inverse problems with a Gaussian prior is considered where the prior scale parameter is estimated using the Empirical Bayes approach. We show that for oversmoothing priors, i.e. when a function is assumed to be a priori smoother that the true unknown function, the posterior distribution of the unknown function with plugged in empirical Bayes estimator of the prior scale parameter achieves the minimax rate of contraction in the considered cases, namely mildly and severely ill-posed inverse problems with true function in Sobolev class, and severely ill-posed problems with analytic true functions. Considered error models are white noise and fractional Brownian motion. We will illustrate behaviour of the empirical Bayes estimator and the empirical Bayes posterior distribution on simulated data.

EO055 Room R23 TRACK: BAYESIAN SEMI- AND NON-PARAMETRIC METHODS II Chair: Antonio Canale

E0261: **A semiparametric mixture model for positive-definite matrices with applications to neuroimaging**
Presenter: **Dipankar Bandyopadhyay**, Virginia Commonwealth University, United States
Co-authors: Brian Reich, Zhou Lan

Studies on diffusion tensor imaging (DTI) quantifies the diffusion of water molecules in a brain voxel, using an estimated 3x3 symmetric positive definite (p.d.) diffusion tensor matrix. Due to the challenges associated with modeling matrix-variate responses, the voxel-level DTI data is usually summarized by univariate quantities, such as the fractional anisotropy (FA). This leads to evident loss of information. Furthermore, DTI analyses often ignore the spatial associationamong neighboring voxels, leading to imprecise estimates. Although the spatial modeling literature is abundant, modeling spatially dependent p.d. matrices is challenging. To mitigate these issues, we propose a matrix-variate Bayesian semiparametric mixture model, where the p.d. matrices are distributed as a mixture of inverse Wishart distributions, with the spatial dependence captured by a Markov model for the mixture component labels. Related Bayesian computing is facilitated by conjugacy results, and the implementation of the double Metropolis-Hastings algorithm. The simulation study shows that the proposed method is more powerful than the non-spatial methods. We also apply our method to investigate the effect of cocaine use on brain microstructure. By extending spatial statistics to matrix-variate data, we contribute to providing a novel and computationally tractable inferential tool for DTI analysis.

E0299: **Compositions of discrete random probabilities for inference on multiple samples**
Presenter: **Giovanni Rebaudo**, Bocconi University, Italy
Co-authors: Augusto Fasano, Antonio Lijoi, Igor Pruenster

Bayesian hierarchical models have proved to be an effective tool when observations are from different populations or studies, since they naturally allow borrowing information across groups, while allowing the subjects in the same group to share the same unknown distribution. We consider models induced by compositions of discrete random probabilities measures, most notably Pitman-Yor processes. Such compositions are well suited to account for both clustering of populations and clustering of observations. We identify an analytical expression of the distribution of the induced random partition, and this allows us to gain a deeper insight into the theoretical properties of the model while deriving predictive distributions and urn schemes. The proposed models can be used as a building block for addressing problems of density estimation, prediction with species sampling data and testing of distributional homogeneity. The theoretical results further lead us to devise novel MCMC sampling schemes whose effectiveness will be discussed through illustrative examples involving simulated and real data.

E0478: **A latent class modeling approach for differentially private synthetic data for contingency tables**
Presenter: **Andres Barrientos**, Florida State University, United States
Co-authors: Michelle Nixon, Jerry Reiter, Aleksandra Slavkovic

An approach is presented to construct differentially private synthetic data for contingency tables. The algorithm achieves privacy by adding noise to selected summary counts, e.g., two-way margins of the contingency table, via the Geometric mechanism. We posit an underlying latent class model for the counts, estimate the parameters of the model based on the noisy counts, and generate synthetic data using the estimated model. This approach allows the agency to create multiple imputations of synthetic data with no additional privacy loss, thereby facilitating estimation of uncertainty in downstream analyses. We illustrate the approach using a subset of the 2016 American Community Survey Public Use Microdata Sets.

E0677: **A class of conjugate priors for multinomial probit models which includes the multivariate normal one**
Presenter: **Augusto Fasano**, Bocconi University, Italy
Co-authors: Daniele Durante

Multinomial probit models are widely-implemented representations which allow both classification and inference by learning changes in vectors of class probabilities with a set of p observed predictors. Although various frequentist methods have been developed for estimation, inference and classification within such a class of models, Bayesian inference is still lagging. This is due to the apparent absence of a tractable class of conjugate priors, that may facilitate posterior inference on the multinomial probit coefficients. Such an issue has motivated increasing efforts toward the development of effective Markov chain Monte Carlo methods. However, state-of-the-art solutions still face severe computational bottlenecks, especially in large p settings. We prove that the entire class of unified skew-normal (SUN) distributions is conjugate to a wide variety of multinomial probit models. We exploit the SUN properties to improve upon state-of-the-art-solutions for posterior inference and classification both in terms of closed-form results for key functionals of interest, and also by developing novel computational methods relying either on i.i.d. samples from the exact posterior or on scalable and accurate variational approximations based on blocked partially-factorized representations. As illustrated in a gastrointestinal lesions application, the magnitude of the improvements relative to current methods is particularly evident when the focus is on large p applications.

EO059 Room R24 ADVANCES IN MONTE CARLO COMPUTATION FOR DATA SCIENCES Chair: Radu Craiu

E0655: **MCMC-driven importance samplers using partial posteriors**
Presenter: **Luca Martino**, Universidad Rey Juan Carlos, Spain

Many applications require the approximation of intractable integrals involving complex posterior distributions. Monte Carlo methods such as Markov Chain Monte Carlo (MCMC) and Importance Sampling (IS) are often employed to approximate these integrals. We propose adaptive IS schemes driven by MCMC chains each one addressing a partial posterior, i.e., a posterior of subsets of data. Partition of the data is commonly used in distributed frameworks, but we do not consider this setting. The goal is to leverage the use of partial posteriors for constructing more efficient importance samplers. Several schemes are discussed: they are improved versions of the so-called Layered Adaptive Importance Sampling (LAIS) algorithm. We also show an application to minibatch selection. Our schemes are validated in very challenging real-world problems, such as exoplanet detection.

E0723: **Free lunches and subsampling MCMC**
Presenter: **Aaron Smith**, University of Ottawa, Canada
Co-authors: Natesh Pillai, James Johndrow

It is widely known that the performance of MCMC algorithms can degrade quite quickly when targeting computationally expensive posterior distributions, including the posteriors associated with any large dataset. This has motivated the search for MCMC variants that scale well for large datasets. One general approach, taken by several research groups, has been to look at only a subsample of the data at every step. We focus on a simple "no-free-lunch" results which provide some basic limits on the performance of many such algorithms. We apply these generic results to realistic statistical problems and proposed algorithms, and also discuss some special examples that can avoid our generic results and provide a free (or at least cheap) lunch.

E0728: **Scaling Monte Carlo inference for state-space models**
Presenter: **Alexander Shestopaloff**, The Alan Turing Institute, United Kingdom

The iterated conditional Sequential Monte Carlo (cSMC) method is a particle MCMC method commonly used for state inference in non-linear, non-Gaussian state-space models. Standard implementations of iterated cSMC provide an efficient way to sample state sequences in low-dimensional state-space models. However, efficiently scaling iterated cSMC methods to perform well in models with a high-dimensional state remains a challenge. One reason for this is the use of a global proposal, without reference to the current state sequence. In high dimensions, such a proposal will typically not be well-matched to the posterior and impede efficient sampling. We will describe a technique to construct efficient proposals in high dimensions that are local relative to the current state sequence. A second obstacle to the scalability of iterated cSMC is not using the entire observed sequence to construct the proposal. We will introduce a principled approach to incorporating all data in the cSMC proposal at time t. By considering several examples, we will demonstrate that both strategies improve the performance of iterated cSMC for state sequence sampling in high-dimensional state-space models.

E0840: **Control variates and unbiased MCMC**
Presenter: **Radu Craiu**, University of Toronto, Canada
Co-authors: Xiao-Li Meng

The recently proposed L-lag coupling for unbiased MCMC calls for a joint celebration by MCMC practitioners and theoreticians. For practitioners, it circumvents the thorny issue of deciding the burn-in period or when to terminate an MCMC iteration and opens the door for safe parallel implementation. For theoreticians, it provides a powerful tool to establish elegant and easily estimable bounds on the exact error of MCMC approximation at any finite number of iteration. A serendipitous observation about the bias correcting term led us to introduce naturally available control variates into the L-lag coupling estimators. In turn, this extension enhances the coupled gains of L-lag coupling, because it results in more efficient unbiased estimators as well as a better bound on the total variation error of MCMC iterations, albeit the gains diminish with the numerical value of L. The theoretical analysis is supported by numerical experiments that show tighter bounds and a gain in efficiency when control variates are introduced.

EO782 Room R25 INSURANCE ANALYTICS Chair: Tim Verdonck

E0219: **Address identification using telematics: An algorithm to identify dwell locations**
Presenter: **Mina Mostoufi**, Allianz Benelux, Belgium
Co-authors: Christopher Grumiau, Solon Pavlioglou, Tim Verdonck

A method is proposed for exploiting the predictive power of a geo-tagged data set as a means of identification of user-relevant points of interest (POI). The proposed methodology is subsequently applied in an insurance context for the automatic identification of a drivers residence address, solely based on his pattern of movements on the map. The analysis is performed on a real-life telematics dataset. We have anonymized the considered dataset to respect privacy regulations. The model performance is evaluated based on an independent batch of the dataset for which the address is known to be correct. The model is capable of predicting the residence postal code of the user with a high level of accuracy, with an f1

score of 0.83. A reliable result of the proposed method could generate benefits beyond the area of fraud, such as general data quality inspections, one-click quotations, and better-targeted marketing.

E0258: **Model selection based on Lorenz and concentration curves, Gini indices and convex order**
Presenter: **Julien Trufin**, Universite libre de Bruxelles, Belgium

In order to determine an appropriate amount of premium, statistical goodness-of-fit criteria must be supplemented with actuarial ones when assessing performance of a given candidate pure premium. Concentration curves and Lorenz curves are shown to provide actuaries with effective tools to evaluate whether a premium is appropriate or to compare two competing alternatives. The idea is to compare the premium income for sub-portfolios gathering low risks (identified aslow by means of the premiums under consideration) to the true one, or equivalently, to the actual losses. Numerical illustrations performed on hypothetical data and real ones demonstrate the usefulness of the proposed approach.

E0278: **The multinomial micro-level reserving model**
Presenter: **Robin Van Oirbeek**, UAntwerp, Belgium
Co-authors: Emmanuel Jordy Menvouta, Jolien Ponnet, Tim Verdonck, Christopher Grumiau

The estimation of the open liabilities or claims reserve is a very important exercise to ensure the day-to-day activities and even the financial viability of an non-life insurance company. Typically, macro-level reserving models such as the Chain Ladder, are used to this end, estimating the claims reserve for the entire portfolio. However, it will be shown how the claims reserve can be estimated on claim-by-claim basis by the use of a micro-level reserving model. The latter models capture the entire lifecycle of a claim by explicitly modelling the underlying time and payment process. Specific to this version of the micro-level reserving model is that both processes, as well as the IBNR or 'Incurred But Not Reported' model are all modeled using separate multinomial regression models. How all these models are tied together, will be discussed.

E0611: **Underwriting fraud prediction based on conditional density estimations**
Presenter: **Felix Vandervorst**, Allianz Benelux, Belgium

Underwriting premium fraud is the risk of adverse data misrepresentation committed with the intent to benefit from an undue lower premium. We propose a novel approach to quantify underwriting premium fraud risk at application time for an insurance pricing model under identifiability of the conditional distribution assumption and availability of non-misrepresented historical quote data. The approach does not require historical premium fraud labels and adapts to change in pricing policy, unlike most supervised and unsupervised approaches to fraud detection. Moreover, our approach can be used to detect outliers next to predicting underwriting fraud and is extensible to multivariate data misrepresentation. We illustrate the approach with motor insurance underwriting data, where the driver identity may be misrepresented to benefit from an undue lower premium.

CO231 Room R02 TOPICS IN STATISTICAL LEARNING AND TIME SERIES ECONOMETRICS Chair: Julia Schaumburg

C0547: **Self-driving score filters**
Presenter: **Marcin Zamojski**, University of Gothenburg, Sweden

A class of approximate score-driven filters is introduced, which is based on automatic differentiation. The agnostic approach requires that a researcher specify a conditional criterion function and that influence functions for the time-varying parameters exist theoretically. We show that in settings where a score model is assumed to be the data-generating process, self-driving filters produce comparable results to analytically derived optimal filters. The small performance loss comes as a trade-off for vastly increased simplicity and implementability. Self-driving filters may be easily implemented in settings where analytical filters are hard or impossible to derive. Therin performance of self-driving filters rivals or improves typical ad-hoc solutions.

C0555: **A persistence-based decomposition of time series: A tale of two spectra**
Presenter: **Maria Grith**, Erasmus University Rotterdam, Netherlands

Two econometric approaches are investigated to model covariance stationary time series that rely on their decomposition in scale-specific components using a Haar-wavelet transform. These components correspond to different levels of aggregation or frequencies of the data. On the one hand, the multiresolution decomposition (MRD) of a time series applies the transform to a time process. On the other hand, the extended Wold decomposition (EWD) applies the transform to the infinite moving-average parameters and innovations of the Wold representation, which leads to orthogonal components. While this property is theoretically appealing, the empirical estimation of the components in the second approach requires the knowledge of the infinite parameter vector. We investigate the restrictions that lead to equivalent classes of scale-specific data-generating processes or the relations between them and propose MRD-based nonparametric estimators for the EWD framework. The estimation methodology is illustrated in two real data studies on the realized volatility and the interactions between macroeconomic variables with persistent components at selected scales.

C0786: **Forecasting heavy-tailed noncausal processes and bubble crash odds**
Presenter: **Sebastien Fries**, Vrije Universiteit Amsterdam, Netherlands

Noncausal or anticipative, heavy-tailed processes generate trajectories featuring locally explosive episodes akin to speculative bubbles in financial time series data. For X_t, a two-sided infinite alpha-stable moving average, conditional moments up to integer-order four are shown to exist provided X_t is anticipative enough, despite the process featuring infinite marginal variance. Formulae of these moments at any forecast horizon under any admissible parameterisation are provided. Under the assumption of errors with regularly varying tails, closed-form formulae of the predictive distribution during explosive bubble episodes are obtained, and expressions of the ex-ante crash odds at any horizon are available. It is found that the noncausal autoregression of order 1 (AR(1)) with AR coefficient ρ and tail exponent α generates bubbles whose survival distributions are geometric with parameter ρ^{α}. This property extends to bubbles with arbitrarily-shaped collapse after the peak, provided the inflation phase is noncausal AR(1)-like. It appears that mixed causal-noncausal processes generate explosive episodes with certain dynamics which could reconcile rational bubbles with tail exponents greater than 1. The use of the conditional moments is illustrated in a bubble-timing portfolio allocation framework, and an application of the closed-form predictive crash odds to the Nasdaq and S&P500 series is provided.

C1113: **Natural capital and the term-structure of sovereign bonds**
Presenter: **Dieter Wang**, VU Amsterdam, Netherlands

Sustainability more broadly and environmental risks, in particular, have become a key issue in the financial world. The aim is to explore the relationship between the country's national wealth, comprising human, natural and produced capital, and its sovereign bonds. We study 22 high and 12 middle-income countries and relate the level, slope and curvature of their government yield curves with their wealth accounts. We document significant and profound impacts of wealth, which reflects a country's long-term growth potential, even after controlling for macro-financial variables and global bond factors. Natural capital raises yields in high-income countries but lowers yields in middle-income countries. Renewable natural wealth, such as forests, agricultural land or protected areas, can be seen as either worthwhile investments or opportunity costs. We discuss the effect of non-renewable fossil fuels and minerals in the context of the natural resource curse and the energy transition.

CO307 Room R03 ADVANCES IN FINANCIAL ECONOMETRICS Chair: Yifan Li

C0427: **Estimating and forecasting long-horizon dollar return skewness**
Presenter: **Jiayu Jin**, The University of Manchester, United Kingdom
Co-authors: Kevin Aretz, Yifan Li

The aim is to develop a parametric estimator of the physical skewness of an assets discrete (i.e. dollar) return over long horizons from the assumption that the assets value can be modelled using a stochastic process from the affine stochastic volatility (ASV) model class. Taking compounding and return dependence effects into account, we demonstrate that our estimator is close to unbiased and efficient, setting it apart from other recent estimators. In a further contrast to those other estimators, it also lends itself naturally to forecasting skewness. Applying our estimator to some representative stock indices, we show that the skewness of long-horizon dollar returns is far less extreme than suggested in the current literature.

C0536: **Liquidity and price informativeness in blockchain-based markets**
Presenter: **Stefan Voigt**, University of Copenhagen, Denmark

Blockchain-based markets impose substantial costs on cross-market trading due to the decentralized and time-consuming settlement process. We quantify the impact of the time-consuming settlement process in the market for Bitcoin on arbitrageurs activity. The estimation rests on a novel threshold error correction model that exploits the notion that arbitrageurs suspend trading activity when arbitrage costs exceed price differences. We estimate substantial arbitrage costs that explain 6% of the observed price differences, where more than 75% of these costs can be attributed to the settlement process. We also find that a 10 bp decrease in latency-related arbitrage costs simultaneously results in a 3 bp increase of the quoted bid-ask spreads. We reconcile this finding in a theoretical model in which liquidity providers set larger spreads to cope with high adverse selection risks imposed by increased arbitrage activity. Consequently, efforts to reduce the latency of blockchain-based settlement might have unintended consequences for liquidity provision. In markets with substantial adverse selection risk, faster settlement may even harm price informativeness.

C0763: **Forecasting realized volatility with echo state networks**
Presenter: **Michael Grebe**, The University of Manchester, United Kingdom

The ability of artificial neural Echo State Networks (ESN) to forecast daily realised volatilities is examnined using high-frequency data. The analysis allows for different asset classes, market conditions and forecasting horizons. ESNs with different architectures were constructed and the forecasting performance compared with the HAR model as a benchmark. The results show that ESNs produce accurate volatility forecasts, where ESNs with feedback connections outperformed those without feedback loops and the parameters were simultaneously optimized using grid searching algorithm to minimize the QLIKE loss function. Across different asset classes, Echo State Networks performed better for equity indices and anti-cyclical stocks with significantly better QLIKE statistics than in the benchmark model, whereas for cyclical stocks the performance was insignificantly better or worse than in HAR. ESN showed better performance for shorter time horizons and in sub-samples with higher volatility persistence but was less sensitive to regime changes such as crisis and pre-crisis periods. Overall, the analysis has proven ESNs as a valid alternative to the HAR model and an easy-to-use and accurate tool with an improvement potential through further meta-parameter optimization.

C0999: **Measuring persistence in volatility spillovers**
Presenter: **Onno Kleen**, Erasmus University Rotterdam, Germany
Co-authors: Christian Conrad, Enzo Weber

Volatility spillovers in multivariate GARCH-type models are analyzed. We show that the cross-effects between the conditional variances determine the persistence of the transmitted volatility innovations. In particular, the effect of a foreign volatility innovation on a conditional variance is even more persistent than the effect of own innovations unless it is offset by an accompanying negative variance spillover of sufficient size. Moreover, ignoring a negative variance spillover causes a downward bias in the estimate of the initial impact of the foreign volatility innovation. Applying the concept to portfolios of small and large firms, we find that shocks to small firm returns affect the large firm conditional variance once we allow for (negative) spillovers between the conditional variances themselves.

CO255 Room R04 MODELLING, FORECASTING, VOLATILITY AND ACCURACY Chair: Mauro Costantini

C0405: **Systemic risk and severe real economy downturns: A sparse meta-analysis**
Presenter: **Michele Costola**, SAFE, Goethe University Frankfurt, Germany
Co-authors: Massimiliano Caporin, Bertrand Maillet, Jean-Charles Garibal

After the major financial crisis of 2008, several systemic risk measures were proposed in the financial literature to quantify the magnitude of financial system distress. We suggest the construction of a novel overall meta-index for the measurement of systemic risk based on a sparse principal component analysis of main systemic risk measures, with the ultimate aim to provide an index with a sound dynamic and proven explicit links to the stress of the financial system and future severe economic recessions.

C0328: **Distilling large information sets to forecast commodity returns: Automatic variable selection or hidden Markov models**
Presenter: **Massimo Guidolin**, Baffi CAREFIN, Italy
Co-authors: Manuela Pedio

The out-of-sample, recursive predictive accuracy is investigated for (fully hedged) commodity future returns of two sets of forecasting models, i.e., hidden Markov chain models (in which the coefficients of predictive regressions follow a regime-switching process) and stepwise variable selection algorithms (in which the coefficients of predictors not selected are set to zero). We perform the analysis under four alternative loss functions, i.e., squared and the absolute value and the realized, portfolio Sharpe ratio and MV utility when the portfolio is built upon optimal weights computed solving a standard MV portfolio problem. We find that neither HMM nor stepwise regressions manage to systematically (or even just frequently) outperform a plain vanilla AR benchmark according to RMSFE or MAFE statistical loss functions. However, in particular, stepwise variable selection methods create economic value in out-of-sample mean-variance portfolio tests. Because we impose transaction costs not only ex-post but also ex-ante, so that an investor uses the forecasts of a model only when they increase expected utility, the economic value improvement is maximum when transaction costs are taken into account.

C0320: **Forecasting directional volatility in the US market**
Presenter: **Eirini Bersimi**, University of Kent, United Kingdom

The aim is to assess the performance of the combination of forecasts for future volatility using measures of directional accuracy (DA) and direction forecast value (DV). Diverse alternative forecast combinations are considered: i) a hierarchical encompassing-MSE procedure; ii) a weighted average forecast method for directional change; iii) two schemes based on DA. An empirical application for the forecasts of daily return volatility for the US and the UK stock markets is carried out.

C0173: **DSGE models with expectations correction and directional forecast accuracy**
Presenter: **Mauro Costantini**, University of L'aquila, Italy

The aim is to investigate the forecasting performance of a small scale New Keynesian model based on expectations correction. The forecast evaluation of the model is conducted using a measure of directional accuracy and directional value. This is the novelty of the paper. A Monte Carlo study is performed under different scenarios using DSGE and non-structural models. An empirical application to UK quarterly data over the period 1986-2019 is also carried out. The main results show that the small scale DSGE with expectations correction is competitive against non-structural models, both in terms of directional accuracy and directional value.

CO305 Room R06 QUANTITATIVE MANAGEMENT Chair: Serge Darolles

C0276: **Futures market liquidity and the trading cost of trend following strategies**
Presenter: **Charles Chevalier**, Universite Paris Dauphine, France
Co-authors: Serge Darolles
A unique dataset reporting the trading of an institutional asset manager implementing trend following strategies is used to estimate the associated transaction costs. With information both at the trade and the fund levels, we disentangle the impact of the execution quality from the management decisions on these costs. We show that the disappointing performances observed for trend following these recent years are explained by a drop in the volatility of the futures markets these strategies generally trade.

C0277: **Do ETFs increase the comovements of their underlying assets? Evidence from a switch in ETF replication technique**
Presenter: **Thomas Marta**, Universite Paris Dauphine, France
Co-authors: Fabrice Riva
The impact of Exchange-Traded Funds (ETFs) on their constituent securities is investigated. The analysis is performed using a novel identification which exploits the switch from synthetic to physical replication of a large French ETF. We find that constituent stocks experience greater commonality, both in returns and in liquidity, after the switch. The effect on return commonality appears stronger for the least liquid stocks included in the ETF. Moreover, we present evidence that ETF arbitrage is the transmission mechanism of the comovements.

C0398: **Factor investing: The missing link between active and passive management**
Presenter: **Wale Dare**, HEC Liege, Belgium
Co-authors: Marie Lambert, Serge Darolles, Guillaume Monarcha
The number of equity indices is more than 70 times greater than all the listed stocks in the world. A subset of these indices, alternative risk premia (ARP), attempt to capture market anomalies, i.e. sources of return not priced by traditional factors. These investment strategies are first back-tested before being commercialized at inception date. An event study on the significance of these alternative risk premia is carried out to explain active management returns pre- and post-inception date. The empirical results show that alternative risk premia significantly explain active management returns pre-inception but that their relevance significantly decreases after inception. The decrease in significance is directly related to the number of indices replicating the same strategy available on the market. The implications of these results are twofold. First, the results support the existence of an economic cycle of alternative risk premia which starts as an active investment strategy to a passive investment package into indices. Secondly, we give evidence of a crowd effect leading to a change in the business model of asset managers from active-based returns to fee-based revenue.

C0839: **ML in asset management**
Presenter: **Rafael Molinero**, Molinero Capital Management, United States
Machine Learning is a hot topic in Asset Management. Most companies in the financial industry are looking into various ML models to see if they can apply them in their business. We will look into the various potential applications of ML in Asset Management, whether trading, allocation, pricing for financial or even physical assets. We will also review the limitations and difficulties met by various ML models when used in Finance, especially as they face low signal to noise ratios as well as multiple regime shifts.

CO067 Room R07 CLIMATE FINANCE Chair: Monica Billio

C0776: **The impact of climate on economic and financial cycles: A Markov-switching panel approach**
Presenter: **Ayokunle Anthony Osuntuyi**, University Ca' Foscari of Venice, Italy
Co-authors: Monica Billio, Roberto Casarin, Malcolm Mistry, Enrica De Cian
The impact of climate shocks on 13 eurozone economies in different phases of business and financial cycle is examined. A Bayesian Panel Markov-switching framework is proposed to jointly estimate the impact of extreme weather events on the economies as well as the interaction between business and financial cycles. Results from the empirical exercise suggest that extreme weather events impact asymmetrically across the different phases of the economy and heterogeneously across the EU countries. A further empirical study indicates that the manufacturing output, a component of the industrial production index, constitutes the main channel through which climate shocks impacts the EU economies.

C0854: **The climate spread of corporate and sovereign bonds**
Presenter: **Stefano Battiston**, CaFoscari University of Venice, Italy
Co-authors: Irene Monasterolo
Climate risk generates a new type of financial risk that standard approaches to risk management are not adequate to handle. Amidst the growing concern about climate change, central banks, financial regulators and policymakers are concerned with the risk of a disorderly low-carbon transition, i.e. a situation in which a sudden introduction of climate policy (e.g. a carbon tax) cannot be fully anticipated by investors and affects large portions of assets, causing asset price volatility (both positive and negative). We develop a model that allows computing the valuation adjustment of corporate and sovereign bonds conditioned to climate transition risk, based on available forward-looking knowledge on climate policy scenarios provided by climate economic models. Our model allows investigating the impact of the endogeneity and deep uncertainty of future scenarios on both the valuation of individual bonds and on standard risk metrics for leveraged investors, considering the role of fossil fuels and carbon-intensive activities in the economy of countries. Based on analytical results complemented by copula simulations, we show that the probability of default of investors is very sensitive to the characteristics of climate policy scenarios, including the default probability of the bonds in the portfolio and their correlation. Thus, Climate stress test exercises need to allow for wide enough sets of scenarios to avoid underestimation of losses.

C0856: **The macroeconomic and financial impact of compounding COVID-19,climate change and financial risks**
Presenter: **Irene Monasterolo**, Vienna University, Austria
The COVID19 crisis has been treated as a public health issue with short-term economic and financial repercussions. This approach neglects the compounding of COVID19 with other main sources of risks in our society, climate change and finance. Neglecting compound risk could lead to underestimating losses, and induce unnecessary trade-offs between investing in the COVID19 recovery or climate change mitigation. We apply the EIRIN Stock Flow Consistent behavioural model to quantitatively assess the direct and indirect impacts of compounding COVID-19, climate and financial risk. The model allows considering the implications of risk uncertainty and system complexity on agents access to information, intertemporal preferences, the formation of expectations and decision making. We apply and calibrate the model on Mexico and analyse the implications of compound risk on GDP, unemployment, private investment, fiscal impacts (e.g., expenditure, revenue, fiscal space) and impacts on households assets and inequality. The results show that risk compounding could amplify losses by triggering reinforcing feedbacks and non-linearities in the economy that in turn lead to hysteresis. Further, timely and targeted public spending plays a main role to mitigate the short term economic impacts of the shocks, by signalling both domestic demand and supply, and affecting agents expectations.

C0661: **COVID-19 spreading in the financial networks**
Presenter: **Monica Billio**, University of Venice, Italy
Co-authors: Roberto Casarin, Michele Costola
Network models represent a useful tool to describe the complex set of financial relationships among heterogeneous firms in the system. A dynamic model is proposed for temporal multilayer networks where the different firms' exposures such as return, volatility, and macroeconomic variables

are represented as layers. Each adjacency matrix of a layer is modeled as a function of the other layers allowing to characterize the relationships according to the selected exposures. In the empirical analysis, we study the topology of the network before and after the spreading of the COVID-19.

CO303 Room R08 UNCERTAINTY IN EMPIRICAL MACROECONOMICS Chair: Alessia Paccagnini

C0305: **An analysis of investments and their drivers in Lithuania**

Presenter: **Mariarosaria Comunale**, Bank of Lithuania, Lithuania

Recent developments in investments in Lithuania using a broad set of possible drivers, including EU funds, are analysed. We apply a Bayesian VAR setup with data from 1995Q1 to 2019Q4. We also look at business vs government investments and different types of investments, especially innovative investments, drawing comparisons. We find that the data on business investments basically drive total investments. The main outcomes are mostly in line with the literature, but we do see some crucial differences across types. Among the key results: 1) we see a minor role for lending rates; 2) we confirm the vital role of demand-side variables (foreign demand or private consumption); 3) there is pro-cyclicality of government investments and a positive correlation with business investments; 4) the uncertainty is key for some sectors and positively drives more innovative/intangible investments and 5) EU funds do feed investments, but we see a crowding out in the short-run for business-related investments, while there are some positive contributions to public investments.

C0407: **Macroeconomic and financial risks: A tale of volatility**

Presenter: **Molin Zhong**, Federal Reserve Board, United States

Co-authors: Chiara Scotti, Dario Caldara

The joint dynamics of the mean and volatility of financial and macroeconomic variables are modeled through a structural vector autoregression with stochastic volatility (SV-VAR). Co-movements in the means and volatilities can produce time-variation and asymmetries in the conditional distributions of the endogenous variables. We first study the evolution of the linkages between macroeconomic and financial tail risks, and the implications of volatility fluctuations for these risks. We then exploit the structure of the model - together with some identification assumptions - to understand the role of level and volatility shocks through a counterfactual experiment. We find that level shocks generate time-varying risk through endogenous volatility, and volatility shocks drive large movements in macroeconomic and financial risk.

C0585: **Identifying uncertainty shock: A Bayesian mixed frequency VAR approach**

Presenter: **Fabio Parla**, Central Bank of Ireland, Ireland

Co-authors: Alessia Paccagnini

The aim is to investigate the transmission of COVID-19-induced financial uncertainty shocks to proxy global financial conditions and real economic activity by extending a previous empirical analysis to a mixed-frequency data sampling (MIDAS) approach. In detail, we proxy global financial uncertainty shocks by using the VIX. Global financial conditions and real economic activity are proxied, respectively, by the global financial cycle index and the world industrial production index. We estimate a Mixed-Frequency Vector Autoregressive model fitted to daily/weekly VIX and to monthly observations on the GFC index and the WIP. The model is estimated over January 1990-April 2019. To account for parameter proliferation, we estimate the model by adopting Bayesian techniques. Overall, the comparison of the impulse responses obtained from the estimation of an MF-VAR with those from a standard (common frequency) VAR suggests moderate evidence of a temporal aggregation bias corroborated by differences in the magnitude of the responses and in the uncertainty around the estimates. These differences are more pronounced when we increase the discrepancy between high- and low-frequency variables (e.g. daily vs monthly data).

C0929: **The distributional impact of the pandemic**

Presenter: **Sinem Hacioglu**, Bank of England, United Kingdom

The top quartile of the income distribution accounts for almost half of the pandemic-related decline in aggregate consumption, with expenditure for this group falling much more than income. In contrast, the bottom quartile of the income distribution has seen the smallest spending cuts and the largest earnings drop. Still, their total incomes have fallen by much less because of the increase in government benefits. The decline in consumers spending preceded the introduction of the lockdown, whose partial lifting has triggered a stronger recovery in sectors with a lower contact rate. The largest spending contractions are concentrated in the most affluent regions. These conclusions are based on detailed high-frequency transaction data on spending, earnings and income from a large Fintech company in the United Kingdom.

Saturday 19.12.2020 15:25 - 17:30 Parallel Session E – CFE-CMStatistics

EO600 Room R11 FUNCTIONAL DATA AND COMPLEX DATA ANALYSIS Chair: Kehui Chen

E0217: **High-dimensional, multiscale online changepoint detection**
Presenter: **Tengyao Wang**, University College London, United Kingdom
Co-authors: Yudong Chen, Richard Samworth

A new method is introduced for high-dimensional, online changepoint detection in settings where a p-variate Gaussian data stream may undergo a change in mean. The procedure works by performing likelihood ratio tests against simple alternatives of different scales in each coordinate, and then aggregating test statistics across scales and coordinates. The algorithm is online in the sense that its worst-case computational complexity per new observation, namely $O(p^2 \log(ep))$, is independent of the number of previous observations; in practice, it may even be significantly faster than this. We prove that the patience, or average run length under the null, of our procedure is at least at the desired nominal level, and provide guarantees on its response delay under the alternative that depend on the sparsity of the vector of mean change. Simulations confirm the practical effectiveness of our proposal, which is implemented in the R package ocd.

E0732: **A wavelet-based independence test for functional data with an application to meg functional connectivity**
Presenter: **Xiaoke Zhang**, George Washington University, United States
Co-authors: Rui Miao, Raymond Ka Wai Wong

Measuring and testing the dependency between multiple random functions is often an important task in functional data analysis. In the literature, a model-based method relies on a model which is subject to the risk of model misspecification. In contrast, a model-free method only provides a correlation measure which is inadequate to test independence. We adopt the Hilbert-Schmidt Independence Criterion (HSIC) to measure the dependency between two random functions. We develop a two-step procedure by first pre-smoothing each function based on its discrete and noisy measurements and then applying the HSIC to recovered functions. To ensure the compatibility between the two steps such that the effect of the pre-smoothing error on the subsequent HSIC is asymptotically negligible, we propose to use wavelet soft-thresholding for pre-smoothing and Besov-norm-induced kernels for HSIC. We also provide the corresponding asymptotic analysis. The superior numerical performance of the proposed method over existing ones is demonstrated in a simulation study. Moreover, in a magnetoencephalography (MEG) data application, the functional connectivity patterns identified by the proposed method are more anatomically interpretable than those by existing methods.

E0710: **Stochastic approximations to optimal transport**
Presenter: **Yoav Zemel**, University of Cambridge, United Kingdom
Co-authors: Axel Munk, Marcel Klatt

Optimal transport is now a popular tool in statistics, machine learning, and data science. A major challenge in applying optimal transport to large-scale problems is its high computational cost. We propose a simple resampling scheme for fast randomized approximate computation of optimal transport distances on finite spaces. This scheme operates on a random subset of the full data and can use any exact algorithm as a black-box back-end, including state-of-the-art solvers and entropically penalized versions. We give non-asymptotic deviation bounds for its accuracy in the case of discrete optimal transport problems. We show that in many important instances, including images (2D-histograms), the approximation error is independent of the size of the full problem. We present numerical experiments demonstrating the excellent approximation that can be obtained while decreasing the computation time by several orders of magnitude. We will also discuss further, recently obtained results on the limiting distribution of the optimal transport plan.

E1031: **Nonparametric estimation of repeated densities with heterogeneous sample sizes**
Presenter: **Xiongtao Dai**, Iowa State University, United States
Co-authors: Jiaming Qiu, Zhengyuan Zhu

The estimation of densities in multiple subpopulations is considered, where the available sample size in each subpopulation greatly varies. For example, in epidemiology, different diseases may share similar pathogenic mechanism but differ in their prevalence. Without specifying a parametric form, the proposed approach pools information from the population and estimate the density in each subpopulation in a data-driven fashion. Low-dimensional approximating density families in the form of exponential families are constructed from the principal modes of variation in the log-densities, within which subpopulation densities are then fitted based on likelihood principles and shrinkage. The proposed methods are shown to be interpretable and efficient in simulation as well as applications to electronic medical record and rainfall data.

E1046: **Testing the equality of distributions of two independent Hilbert-valued random elements**
Presenter: **Gil Gonzalez-Rodriguez**, University of Oviedo, Spain
Co-authors: Ana Colubi, Wenceslao Gonzalez-Manteiga, Manuel Febrero-Bande

The aim is to develop a bootstrap test to check whether two independent random elements taking on values in a separable Hilbert space have the same distribution or not. A transformation of both random elements into a new separable Hilbert space will be considered in such a way that the equality of expectations of the transformed random elements characterizes the equality of distributions. Consequently, any bootstrap procedure to check the equality of means of two independent random elements taking on values in a separable Hilbert space can be used to solve the original problem. This simple strategy is complemented by checking that the bootstrap procedure can be used without the need to transform the random elements, by applying some simple transformations in the original space.

EO277 Room R12 RECENT DEVELOPMENTS ON ROBUST FUNCTIONAL DATA ANALYSIS Chair: Sara Lopez Pintado

E0452: **The BRIk and FABRIk algorithms for improving k-means clustering recovery**
Presenter: **Aurora Torrente Orihuela**, Universidad Carlos III de Madrid, Spain
Co-authors: Javier Albert Smet, Juan Romo

The k-means algorithm is widely used in various research fields because of its fast convergence to the cost function minima; however, it frequently gets stuck in local optima as it is sensitive to initial conditions. The BRIk algorithm is a simple, computationally feasible and efficient method that provides k-means with a set of initial seeds to cluster datasets of arbitrary dimensions. In terms of clustering recovery, it drastically improves k-means results with respect to other widely-used initialization procedures. It relies on clustering a set of tighter (thus easier to separate) centroids derived from bootstrap replicates of the data and on the use of the versatile Modified Band Depth to identify the deepest point of each cluster. On the other hand, FABRIk is a recent functional-data extension of the BRIk algorithm for longitudinal data, where appropriate B-splines are fit to the observations and a resampling process is used to handle issues such as noise or missing data. When run with simulated and real data sets, FABRIk outperforms the alternative techniques, including BRIk and functional-data versions of its competitors.

E0853: **Uncertainty analysis of contagious processes based on a functional approach**
Presenter: **Zonghui Yao**, Northeastern University, United States
Co-authors: Dunia Lopez-Pintado, Sara Lopez Pintado

One of the most intricate and important phenomena studied by sociologists, economists, and epidemiologists are predicting the contagion process of a new idea, product, or disease in a population characterized by a complex network of interactions. All of these phenomena have in common that

the unit of study is a contagion curve (proportion of "infected" over time). This curve is typically very complicated to anticipate given the stochastic and uncertain nature of the adoption dynamics. The aim is to measure the variability formally and, ultimately, the unpredictability of the contagion process using novel statistical methods based on functional data depth and systematic large-scale simulation studies. Starting with the well-known Susceptible-Infected-Susceptible (SIS) model and an Erdos-Reny random network, we show that the unpredictability of the process is related to intermediate network density and disease infectivity. The analysis can be extended to a more advanced Susceptible-Infected-Immune-Death (SIID) model and a more realistic network with intervention. The main contribution is to use functional data tools to study the uncertainty of contagion processes on different networks.

E0874: **Depth in function spaces: Approaches and challenges**
Presenter: **Stanislav Nagy**, Charles University, Czech Republic

The depth is a statistical tool that introduces an ordering of points in multivariate/function spaces. Some of the natural properties a reasonable depth should satisfy in finite-dimensional spaces however lose appeal as the dimension grows. We discuss the depth in finite-dimensional spaces and outline particular difficulties when attempting to generalize depths to functional, or other infinite-dimensional data.

E0795: **Group selection with Bayesian high dimensional modeling**
Presenter: **Naveen Naidu Narisetty**, University of Illinois at Urbana-Champaign, United States

In many applications with high-dimensional covariates, including functional covariates, the covariates are naturally structured into different groups which can be used to perform efficient statistical inference. We propose a Bayesian hierarchical model with a spike and slab prior specification to perform group selection in high-dimensional linear regression models. While several penalization methods and more recently, some Bayesian approaches are proposed for group selection, the theoretical properties of Bayesian approaches have not been studied extensively. We provide novel theoretical results for group selection consistency under spike and slab priors which demonstrate that the proposed Bayesian approach has advantages compared to penalization approaches. Our theoretical results accommodate flexible conditions on the design matrix and can be applied to commonly used statistical models such as functional additive models. A shotgun stochastic search algorithm is adopted for the implementation of our proposed approach. We illustrate through simulation studies that the proposed method has better performance for group selection compared to a variety of existing methods.

E0646: **Localization processes for functional data analysis**
Presenter: **Raul Jimenez**, Universidad Carlos III de Madrid, Spain
Co-authors: Antonio Elias, Joseph Yukich

An alternative is proposed to k-nearest neighbors for functional data whereby the approximating neighbor curves are piecewise functions built from a functional sample. Instead of a distance on a function space we use a locally defined distance function that satisfies stabilization criteria. We exploit this feature to develop the asymptotic theory when the number of curves is large enough or when a finite number of curves is observed at time-points coinciding with the realization of a point process with intensity increasing to infinity. We use these results to investigate the problem of estimating unobserved segments of a partially observed functional data sample as well as to study the problem of functional classification and outlier detection. For these problems, we discuss methods that are competitive with and often superior to benchmark predictions in the field.

EO139 Room R13 RECENT ADVANCES IN STATISTICS OF EXTREMES Chair: Raphael Huser

E0371: **On the global optimality and asymptotic posterior normality of the generalized extreme value likelihood function**
Presenter: **Likun Zhang**, Lawrence Berkeley National Lab, United States

The three-parameter generalized extreme value (GEV) distribution arises from classical univariate extreme value theory and is in common use for analyzing the far tail of observed phenomena. Curiously, important asymptotic properties of likelihood-based estimation under this standard model have yet to be established. We first show that the maximum likelihood estimator is global and unique. We then use these results to demonstrate the asymptotic normality of the corresponding posterior distribution under regular priors. These properties are crucial to performing Bayesian optimal derivations such as optimal decision rules and reference prior distributions under the GEV distribution.

E0429: **Modelling the tail behaviour of precipitation aggregates using conditional spatial extremes**
Presenter: **Jordan Richards**, Lancaster University, United Kingdom
Co-authors: Jonathan Tawn

Fluvial flooding is not caused by high-intensity rainfall at a single location; rather, it is caused by the extremes of precipitation events aggregated over spatial catchment areas. Accurate modelling of the tail behaviour of such events can help to mitigate the financial aspects associated with floods, especially if river defences are built within specification to withstand an $n-$year event of this kind. Within an extreme value analysis framework, univariate methods for estimating the size of these $n-$year events are well studied and cemented in asymptotic theory. To complement these techniques, we develop a high-resolution spatial model for extreme precipitation by providing a fully spatial extension of the conditional approach for modelling multivariate extremes. We simulate realistic precipitation fields from this model and use univariate techniques to make inference about the extremal behaviour of aggregates over specified spatial domains. The challenge of zero precipitation data is overcome, and further applications of the model are discussed. The model is fit to data from a convection-permitting forecast model within the 2018 UK Climate Projections (UKCP18).

E0620: **Spatial hierarchical modeling of threshold exceedances using rate mixtures**
Presenter: **Rishikesh Yadav**, King Abdullah University of Science and Technology, Saudi Arabia
Co-authors: Raphael Huser, Thomas Opitz

New flexible univariate tail models are developed for light-tailed and heavy-tailed data, which extend a hierarchical representation of the generalized Pareto (GP) limit for univariate threshold exceedances. These models can accommodate departure from asymptotic threshold stability in finite samples while keeping the asymptotic GP distribution as a special (or boundary) case and can be used to model the tails and the bulk regions jointly without losing much flexibility. Spatial dependence is modeled through a latent process, while the data are assumed to be conditionally independent. We design penalized complexity priors for crucial model parameters to shrink toward a simpler reference GP distribution with moderately heavy tails. We fit our models in fairly high dimensions based on Markov chain Monte Carlo by exploiting the Metropolis-adjusted Langevin algorithm (MALA), which guarantees fast convergence of Markov chains using efficient block proposals for the latent variables. We also develop an adaptive scheme to calibrate the MALA tuning parameters. Furthermore, our models avoid the expensive numerical evaluations of multifold integrals in censored likelihood expressions. We demonstrate our new methodology by simulation and application to a dataset of extreme rainfall episodes that occurred in Germany. We will extend this framework to the modeling of spatio-temporal extremes.

E0656: **Conex-Connect: Learning patterns in extremal brain connectivity 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 brain disease affecting more than 50 million people globally. An epileptic seizure occurs as an abnormal temporary shock to the neuronal system, with results varying from a very brief and almost imperceptible loss of consciousness to uncontrollable spasms. Epilepsy is frequently diagnosed with electroencephalograms (EEGs), and statistical methods are widely used to analyze EEG signals. We propose a new approach to characterize brain connectivity during an epileptic seizure. Our method models the conditional extremal dependence for brain

connectivity (Conex-Connect). It is a pioneering method in linking the association between extreme values of higher oscillations at a reference channel with the other channels of the brain network. We applied our method to EEG data from a patient diagnosed with left temporal lobe epilepsy, revealing changes in the conditional extremal dependence of brain connectivity. Pre-seizure, the dependence is notably stable for all channels when conditioning on extreme values of the focal seizure area. Post-seizure, the dependence between channels is weaker, and dependence patterns are more "chaotic". Also, in terms of spectral decomposition, high values of the higher frequency band are the most relevant features to explain the conditional extremal dependence of brain connectivity.

E0660: **Semiparametric bivariate modelling with flexible extremal dependence**
Presenter: **Manuele Leonelli**, IE University, Spain

Inference over multivariate tails often requires several assumptions which may affect the assessment of the extreme dependence structure. Models are usually constructed in such a way that extreme components can either be asymptotically dependent or be independent of each other. Recently, there has been an increasing interest in modelling multivariate extremes more flexibly by allowing models to bridge both asymptotic dependence regimes. In this talk, Novel semiparametric approaches are discussed which allow for a variety of dependence patterns, be them extremal or not, by using in a model-based fashion the full dataset. These build on previous work for inference on marginal exceedances over a high, unknown threshold, by combining it with flexible, semiparametric copula specifications to investigate extreme dependence, thus separately modelling marginals and dependence structure. Because of the generality of the approach, bivariate problems are investigated due to computational challenges, but multivariate extensions are readily available. Empirical results suggest that the proposed approaches can provide sound uncertainty statements about the possibility of asymptotic independence. Estimation of functions of interest for extremes is performed via MCMC algorithms. Environmental applications are used to illustrate the methodology.

EO632 Room R14 RECENT ADVANCES TOWARD UNDERSTANDING DEEP LEARNING Chair: Weijie Su

E0295: **On lower bounds for the bias-variance trade-off**
Presenter: **Alexis Derumigny**, University of Twente, Netherlands
Co-authors: Johannes Schmidt-Hieber

It is a common phenomenon that for high-dimensional and nonparametric statistical models, rate-optimal estimators balance squared bias and variance. Although this balancing is widely observed, little is known whether methods exist that could avoid the trade-off between bias and variance. We propose a general strategy to obtain lower bounds on the variance of an estimator with bias smaller than a prespecified bound. This shows to which extent the bias-variance trade-off is unavoidable and allows the quantification of the loss of performance for methods that do not obey it. The approach is based on some abstract lower bounds for the variance involving the change of expectation with respect to different probability measures as well as information measures such as the Kullback-Leibler or chi-square divergence. In a second part of the article, the abstract lower bounds are applied to several statistical models including the Gaussian white noise model, a boundary estimation problem, the Gaussian sequence model and the high-dimensional linear regression model. For the trade-off between integrated squared bias and integrated variance in the Gaussian white noise model, we propose to combine the general strategy for lower bounds with a reduction technique. This allows us to reduce the original problem to a lower bound on the bias-variance trade-off for estimators with additional symmetry properties in a simpler statistical model.

E0970: **Good classifiers are abundant in the interpolating regime**
Presenter: **Jason Klusowski**, Princeton University, United States
Co-authors: Ryan Theisen, Michael Mahoney

Within the machine learning community, the widely-used uniform convergence framework seeks to answer the question of how complex, over-parameterized models can generalize well to new data. This approach bounds the test error of the *worst-case* model one could have fit to the data, which presents fundamental limitations. Inspired by the statistical mechanics' approach to learning, we formally define and develop a methodology to precisely compute the full distribution of test errors among interpolating classifiers from several model classes. We apply our method to compute this distribution for several real and synthetic datasets with both linear and random feature classification models. We find that test errors tend to concentrate around a small *typical* value ε^*, which deviates substantially from the test error of the worst-case interpolating model on the same datasets, indicating that 'bad' classifiers are extremely rare. We provide theoretical results in a simple setting in which we characterize the full (asymptotic) distribution of test errors, and show that these indeed concentrate around a value ε^*, which we also identify exactly. Our results show that the usual style of analysis in statistical learning theory may not be fine-grained enough to capture the good generalization performance observed in practice, and that approaches based on the statistical mechanics of learning may offer a promising alternative.

E1071: **Local elasticity: A phenomenological approach toward understanding deep learning**
Presenter: **Weijie Su**, The Wharton School, University of Pennsylvania, United States

Motivated by the iterative nature of training neural networks, the following question arises: If the weights of a neural network are updated using the induced gradient on an image of a tiger, how does this update impact the prediction of the neural network at another image (say, an image of another tiger, a cat, or a plane)? To address this question, we will introduce a phenomenon termed local elasticity. Roughly speaking, our experiments show that modern deep neural networks are locally elastic in the sense that the change in prediction is likely to be most significant at another tiger and least significant at a plane, at late stages of the training process. We will illustrate some implications of local elasticity by relating it to the neural tangent kernel and improving on the generalization bound for uniform stability. Moreover, we will introduce a phenomenological model for simulating neural networks, which suggests that local elasticity may arise from the sharing of low-level and intermediate-level features. Finally, we will offer a local-elasticity-focused agenda for future research toward a theoretical foundation for deep learning.

E1157: **An information-geometric distance on the space of tasks**
Presenter: **Pratik Chaudhari**, University of Pennsylvania, United States
Co-authors: Yansong Gao

A distance is computed between tasks modeled as joint distributions on data and labels. We develop a stochastic process that transports the marginal on the data of the source task to that of the target task and simultaneously updates the weights of a classifier initialized on the source task to track this evolving data distribution. The distance between two tasks is defined to be the shortest path on the Riemannian manifold of the conditional distribution of labels given data as the weights evolve. We derive connections of this distance with Rademacher complexity-based generalization bounds; distance between tasks computed using our method can be interpreted as the trajectory in weight space that keeps the generalization gap constant as the task distribution changes from the source to the target. Experiments on image classification datasets show that this task distance helps predict the performance of transfer learning: fine-tuning techniques have an easier time transferring to tasks that are close to each other under our distance.

E1164: **Provable training of certain finite size neural nets at depth 2**
Presenter: **Anirbit Mukherjee**, University of Pennsylvania, United States
Co-authors: Ramchandran Muthukumar

One of the paramount mathematical mysteries is to be able to explain the phenomenon of deep-learning. Neural nets can be made to paint while imitating classical art styles or play chess better than any machine or human ever, and they seem to be the closest we have ever come to achieving "artificial intelligence". But trying to reason about these successes quickly lands us into a plethora of extremely challenging mathematical questions

- typically about discrete stochastic processes. Some of these questions remain unsolved for even the smallest neural nets! We will give a brief introduction to neural nets and describe our recent work about provable training of finitely large depth 2 single filter generalized convolutional nets. Firstly, we will explain how under certain structural and mild distributional conditions our iterative algorithms like “Neuro-Tron" which do not use a gradient oracle can often be proven to train nets using as much time/sample complexity as expected from gradient-based methods but in regimes where usual algorithms like (S)GD remain unproven. Our theorems include the particularly challenging regime of non-realizable data. Secondly, we will explain how for a single ReLU gate slight modification to SGD can get us data-poisoning resilient training.

EO173 Room R15 TRENDS IN THE ANALYSIS OF LARGE AND COMPLEX DATA Chair: Johannes Lederer

E0460: **Scalable estimation of random graph models with dependent edges and parameter vectors of increasing dimension**
Presenter: **Michael Schweinberger**, Department of Statistics, Rice University, United States
Co-authors: Jonathan Stewart
An important question in statistical network analysis is how to construct and estimate models of dependent network data without sacrificing computational scalability and statistical guarantees. We demonstrate that scalable estimation of random graph models with dependent edges is possible, by establishing the first consistency results and convergence rates for maximum pseudo-likelihood estimators for parameter vectors of increasing dimension based on a single observation of dependent random variables. The main results cover models of dependent random variables with countable sample spaces. They may be of independent interest. To showcase consistency results and convergence rates, we introduce a novel class of generalized beta-models with dependent edges and parameter vectors of increasing dimension. We establish consistency results and convergence rates for maximum pseudo-likelihood estimators of generalized beta-models with dependent edges, in dense- and sparse-graph settings.

E0515: **Large-p variable selection in two-stage models**
Presenter: **Haim Bar**, University of Connecticut, United States
Model selection in the large-p small-n scenario is discussed in the framework of two-stage models. Two specific models are considered, namely, two-stage least squares (TSLS) involving instrumental variables (IVs), and mediation models. In both cases, the number of putative variables (either instruments or mediators) is large, but only a small subset should be included in the two-stage model. We use two variable selection methods which are designed for high-dimensional settings, and compare their performance in terms of their ability to find the true IVs or mediators. Our approach is demonstrated via simulations and case studies.

E0688: **Depth normalization of small RNA sequencing: Using data and biology to select the best method**
Presenter: **Li-Xuan Qin**, Memorial Sloan Kettering Cancer Center, United States
Co-authors: Yannick Duren, Johannes Lederer
Deep sequencing has become the most popular tool for transcriptome profiling in cancer research and biomarker studies. Similar to other high through-put profiling technologies such as microarrays, sequencing also suffers from systematic non-biological artefacts that arise from variations in experimental handling. A critical first step in sequencing data analysis is to normalize sequencing depth so that the data can be compared across the samples. A plethora of analytic methods for depth normalization has been proposed, and different normalization methods may lead to different analysis results with no method found to work systematically best. Currently, it is often up to the data analyst to choose a method based on personal preference and convenience. We developed a data-driven and biology-motivated approach to more objectively guide the selection of a depth normalization method for the data at hand. We assessed the performance of this approach using a unique pair of data sets for the same set of tumour samples that were collected at Memorial Sloan Kettering Cancer Center and applied it to additional data sets from the Cancer Genome Atlas for further demonstration.

E0708: **Multiscale geometric feature-extraction for high-dimensional and non-euclidean data**
Presenter: **Wolfgang Polonik**, University of California at Davis, United States
Co-authors: Gabriel Chandler
A method for extracting multiscale geometric features from a data cloud is presented. Each pair of data points is mapped into a real-valued feature function, whose construction is based on geometric considerations. The collection of these feature functions is then being used for further data analysis. Applications include classification, anomaly detection and data visualization. In contrast to the popular kernel trick, the construction of the feature functions is based on geometric considerations. The performance of the methodology is illustrated through applications to real data sets, and some theoretical guarantees are presented.

E0709: **Debiased inverse propensity score weighted for estimation of average treatment effects in high-dimensions**
Presenter: **Rajen D Shah**, University of Cambridge, United Kingdom
Co-authors: Yuhao Wang
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. 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. So, for example, the variances of the responses may be estimated. We show how confidence intervals centred on our estimates may be constructed, and also extend the method to estimate heterogeneous treatment effects.

EO247 Room R16 RECENT ADVANCES IN NON-GAUSSIAN STOCHASTIC PROCESSES Chair: Anastassia Baxevani

E0910: **Extremal clustering under moderate long range dependence and moderately heavy tails**
Presenter: **Gennady Samorodnitsky**, Cornell University, United States
The purpose is to study the clustering of the extremes in a stationary sequence with subexponential tails in the maximum domain of attraction of the Gumbel. We obtain functional limit theorems in the space of random sup-measures and in the space D(0,infty). The limits have the Gumbel distribution if the memory is only moderately long. However, as our results demonstrate rather strikingly, the“heuristic of a single big jump” could fail even in a moderately long-range dependence setting. As the tails become lighter, the extremal behavior of a stationary process may depend on multiple large values of the driving noise.

E0945: **Slepian models for wave asymmetry and particle orbits in Gauss-Lagrange ocean waves**
Presenter: **Georg Lindgren**, Lund university, Sweden
Co-authors: Marc Prevosto
The basic elements in a Gauss-Lagrange model are the water particle orbits. These orbits determine the apparent wave shapes. There are very few detailed observations of the relation between particle orbit orientation and wave shape in the open ocean, and most studies are combinations of theoretical models, wave tank experiments and field data. We present Slepian models for the vertical and horizontal movements in space and time of the water particles that are located at the crests of the space waves at the time of observation. These models contain one regression term, depending on the wave height and curvature at the maximum, and one Gaussian residual part independent of these variables. These Slepian models

are easy to simulate once the rather complicated space-time correlation between vertical and horizontal movements are constructed. We use the models to illustrate the statistical relation between wave asymmetry and particle orbit orientation. We also relate the orientation to the velocities of the top particle. The results are of interest in stochastic fluid dynamics.

E1102: **The Greenwood statistic, stochastic dominance, clustering and heavy tails**
Presenter: **Anna Panorska**, University of Nevada, United States
Co-authors: Tomasz Kozubowski, Marek Arendarczyk

The Greenwood statistic T_n and its functions, including sample coefficient of variation, often arise in testing exponentiality or detecting clustering or heterogeneity. We provide a general result describing the stochastic behavior of T_n in response to stochastic behavior of the sample data. Our result provides a rigorous base for constructing tests and assuring that confidence regions are actually intervals for the tail parameter of many power-tail distributions. We also present a result explaining the connection between clustering and heaviness of tail for several standard classes of distributions and argue its extension to general heavy-tailed families. The results provide theoretical justification for T_n being an effective and commonly used statistic discriminating between regularity/uniformity and clustering in the presence of heavy tails in applied sciences. We also note that the use of Greenwood statistic as a measure of heterogeneity or clustering is limited to data with large outliers, as opposed to those close to zero.

E1107: **Model selection using predictive distributions of stochastic processes**
Presenter: **Jonas Wallin**, Lund University, Sweden

Often when doing forecasting on real data, there is a range of possible models to use. Typically the ranking of the models is determined by the forecasting ability of the models, which in term determined by some scoring function. If this scoring function will, in the long run, select the true model, if available, is known as a proper scoring rule. This scoring function has a long history in the forecasting literature, especially in weather forecasting. If the distribution for one's observations has varying scaling, we show that it is important to take into consideration the result of the scoring rule. This is a characteristic that is overlooked in the scoring rule literature. Further, many of the popular scoring rules, like mse, mae and CRPS do not take scaling into account. We develop a set of new scoring rules and show that these rules take into account the scaling of the predictive distributions.

E1144: **A functional gamma autoregressive processes**
Presenter: **Tomasz Kozubowski**, University of Nevada Reno, United States
Co-authors: Anastassia Baxevani, Krzysztof Podgorski

The purpose is to develop a functional relation of auto-regressive type for gamma Levy process, leading to a new type of stationary functional time series with gamma motions as its marginal distributions. We discuss the basic properties of this construction and provide links to related models. In particular, we describe a new class of Wright Levy processes, which plays an important role in this development.

EO484 Room R17 TOPICS IN HIGH-DIMENSIONAL STATISTICAL INFERENCE Chair: Olga Klopp

E0923: **Online change-point detection in Gaussian graphical models**
Presenter: **George Michailidis**, University of Florida, United States

Piecewise stationary graphical models represent a versatile class for modelling time-varying networks arising in diverse application areas. There is little work in identifying changes in the topology of the network, despite its high relevance to applications. A novel scalable online algorithm is introduced for detecting an unknown number of abrupt changes in sparse Gaussian graphical models with a small delay. The proposed algorithm is based upon monitoring the conditional log-likelihood of all nodes in the network. It can be extended to a large class of continuous and discrete graphical models. Numerical work on both synthetic and real data illustrates the performance of the method.

E1179: **Statistical guarantees for generative models without domination**
Presenter: **Arnak Dalayan**, CREST, ENSAE, IP Paris, France
Co-authors: Nicolas Schreuder, Victor-Emmanuel Brunel

A convenient framework is introduced for studying (adversarial) generative models from a statistical perspective. It consists in modeling the generative device as a smooth transformation of the unit hypercube of a dimension that is much smaller than that of the ambient space and measuring the quality of the generative model through an integral probability metric. In the particular case of an integral probability metric defined through a smoothness class, we establish a risk bound quantifying the role of various parameters. In particular, it clearly shows the impact of dimension reduction on the error of the generative model.

E1055: **Learning with differentiable perturbed optimizers**
Presenter: **Quentin Berthet**, Google Research, France

Machine learning pipelines often rely on optimization procedures to make discrete decisions (e.g., sorting, picking closest neighbors, or shortest paths). Although these discrete decisions are easily computed, they break the back-propagation of computational graphs. In order to expand the scope of learning problems that can be solved in an end-to-end fashion, we propose a systematic method to transform optimizers into operations that are differentiable and never locally constant. The approach relies on stochastically perturbed optimizers and can be used readily together with existing solvers. Their derivatives can be evaluated efficiently, and smoothness tuned via the chosen noise amplitude. We also show how this framework can be connected to a family of losses developed in structured prediction, and give theoretical guarantees for their use in learning tasks. We demonstrate the performance of our approach on various tasks experimentally.

E0329: **Low-rank methods for multi-source, heterogeneous and incomplete data**
Presenter: **Genevieve Robin**, CNRS, France
Co-authors: Julie Josse, Eric Moulines, Robert Tibshirani, Olga Klopp

In modern applications of statistics and machine learning, the urge to collect large data sets often leads to relaxing acquisition procedures and compounding diverse sources. As a result, analysts are confronted with many data imperfections. In particular, data are often heterogeneous, i.e. combine quantitative and qualitative information, incomplete, with missing values caused by machine failures or by the nonresponse phenomenon, and multi-source, when the data result from the aggregation of several data sets. We will present a general framework based on heterogeneous exponential family low-rank models, to analyse heterogeneous, multi-source and incomplete data sets. The theoretical results demonstrate that the method is simultaneously statistically sound with minimax optimal estimation properties and computationally efficient. We will illustrate the empirical behaviour of the method with the analysis of North-African waterbirds monitoring data set.

E0877: **Detecting change points in dynamic networks**
Presenter: **Olga Klopp**, Essec Business School, France
Co-authors: Farida Enikeeva

Changes occur in dynamic networks quite frequently and its detection is an important question in many situations such as fraud detection or cybersecurity. Real-life networks are often incompletely observed due to individual non-response or networks size. We consider the problem of change-point detection at a time sequence of partially observed networks. The goal is to test whether there is a change in the parameters of the network. Our approach is based on the CUSUM test statistic and allows growing size of networks. We show that the proposed test is minimax

optimal and robust to missing links.

EO492 Room R18 COPULAS AND DEPENDENCE MODELLING Chair: Piotr Jaworski

E0336: **Certain copula transformations: From 2D to 3D**
Presenter: **Martynas Manstavicius**, Vilnius University, Lithuania
Co-authors: Gediminas Bagdonas, Egle Gutauskaite
Necessary and sufficient conditions on f for the function $H_f(C)(x,y) = C(x,y)f(1-x-y+C(x,y))$, $x,y \in [0,1]$ to be a bivariate copula for *any* bivariate copula C are known. If, on the other hand, $C(x,y) = \Pi(x,y) = xy$ is fixed, then some of those conditions become no longer necessary, that is, the class of allowable functions f can be substantially enlarged. We have previously determined sufficient conditions on f for $C_f(x,y) = xyf((1-x)(1-y))$ to be a bivariate copula. We will focus on the necessary conditions, implications of such a transformation to probabilistic properties, as well as possibility to extend earlier results to trivariate copulas.

E0465: **On left truncation invariant copulas with fixed Kendall tau**
Presenter: **Piotr Jaworski**, University of Warsaw, Poland
The concordance measures, like for example Kendall tau, are the main numerical characterizations of bivariate copulas. In case of copulas invariant with respect to left truncation (conditioning) of its first representer (LTI), the concordance ordering is fully determined in terms of ordinary differential equations. Basing on this we provide the formulas for the copulas being the lower and upper bounds for LTI copulas with given Kendal tau.

E0468: **Asymptotic behavior of an intrinsic rank-based estimator of the Pickands dependence function constructed from B-splines**
Presenter: **Christian Genest**, McGill University, Canada
A bivariate extreme-value copula is characterized by its Pickands dependence function, i.e., a convex function defined on the unit interval satisfying boundary conditions. This function has been the object of intense study over the past thirty years, and several rank-based estimators thereof are available in the literature. The large-sample behavior of one such estimator is discussed. This estimator is constructed from B-splines and has the advantage of being intrinsic. Under the assumption that the Pickands dependence function is a linear combination of B-splines of order 3 or 4 with a given set of knots, the vector of coefficients is estimated by a constrained penalized minimum distance method. The asymptotic behavior of this estimator will be presented and used to deduce the weak limit of the resulting Pickands dependence function, spectral distribution, and spectral density.

E0434: **The set of zeros for a given copula**
Presenter: **Enrique de Amo**, University of Almeria, Spain
Co-authors: Juan Fernandez Sanchez, Manuel Ubeda Flores
Sufficient and necessary conditions are provided for a given set to be the zero-set for a suitable given copula. The role of k-negations is exploited in this target. We also study several topological properties and the lattice-theoretic structure, and characterize the zero-sets of the class of Archimedean copulas.

E1022: **On extreme biconic and semilinear copulas**
Presenter: **Fabrizio Durante**, University of Salento, Italy
Co-authors: Juan Fernandez Sanchez, Manuel Ubeda Flores
Several optimization problems arising in the study of copula-based stochastic models have motivated the consideration of the extreme points (in the Krein-Milman sense) in the class of copulas. We focus on the extreme elements in the class of semilinear and biconic copulas and provide their characterization. In both cases, a crucial role is played by the diagonal section of the copula, which contains some relevant information about the associated tail dependence.

EO369 Room R19 RECENT DEVELOPMENT IN NETWORK ANALYSIS AND CLUSTER ANALYSIS Chair: Anderson Ye Zhang

E1098: **On the theoretical properties of the network jackknife**
Presenter: **Purnamrita Sarkar**, U. T. Austin, United States
The properties of a leave-node-out jackknife procedure for network data are studied. Under the sparse graphon model, we prove an Efron-Stein-type inequality, showing that the network jackknife leads to conservative estimates of the variance (in expectation) for any network functional that is invariant to node permutation. For a general class of count functionals, we also establish consistency of the network jackknife. We complement our theoretical analysis with a range of simulated and real-data examples and show that the network jackknife offers competitive performance in cases where other resampling methods are known to be valid. In fact, for several network statistics, we see that the jackknife provides more accurate inferences compared to related methods such as subsampling.

E0607: **Inference of high-dimensional modified Poisson-type graphical models**
Presenter: **Zhao Ren**, University of Pittsburgh, United States
The high dimensional graphical model has attracted great attention in biological network analysis with different types of omics data. To tailor the network analysis of important count-valued omics data, various Poisson-type graphical models were proposed, but the statistical inference of these models is not well studied. We investigate statistical inference of each edge for large modified Poisson-type graphical models. The key role in most existing inferential methods is played by a linear projection method to de-bias an initial regularized estimator. The major drawback of this approach in those non-Gaussian graphical models is that an extra sparsity assumption on the linear projection coefficient is required, which cannot be checked in practice. To solve this challenge, we first propose a novel estimator of each edge for Ising model via quadratic programming and show that our estimator is asymptotically normal without the above mentioned extra sparsity condition. In addition, we further show that whenever the extra sparsity condition is satisfied, our estimator is adaptively efficient and achieves the Fisher information. We then extend our approach to modified Poisson-type graphical models. The practical merit of the proposed method is demonstrated by an application to a novel RNA-seq gene expression data set in childhood atopic asthma in Puerto Ricans.

E0900: **Two-sample testing on latent distance graphs with unknown link functions**
Presenter: **Minh Tang**, North Carolina State University, United States
Co-authors: Soumendra Lahiri, Yiran Wang
A valid and consistent test is proposed for the hypothesis that two latent distance random graphs on the same vertex set have the same generating latent positions, up to some unidentifiable similarity transformations. The test statistic is based on first estimating the edge probabilities matrices by truncating the singular value decompositions of the averaged adjacency matrices in each population and then computing a Spearman rank correlation coefficient between these estimates. Experimental results on simulated data indicate that the test procedure has power even when there is only one sample from each population, provided that the number of vertices is not too small. Application on a dataset of neural connectome graphs showed that we can distinguish between scans from different age groups while application on a dataset of epileptogenic recordings.

E0739: **Hierarchical clustering via spectral methods in networks**
Presenter: **Xiaodong Li**, UC Davis, United States

Although spectral clustering has been extensively studied in network analysis, some important issues, such as hierarchical clustering via eigenvectors and determining the number of communities via eigenvalues, have been far less investigated thus far. The theoretical analysis of hierarchical community detection is considered. We show that the graph-Laplacian based spectral hierarchical clustering is consistent under general tree structures and broad ranges of connectivity probabilities. The analysis relies on careful exploitation of the algebraic properties of graph Laplacian and statistical properties of hierarchical SBM.

E0532: **Computationally efficient sparse clustering**
Presenter: **Matthias Loeffler**, ETH Zurich, Switzerland
Co-authors: Alexander Wein, Afonso Bandeira

Statistical and computational limits of clustering are studied when the means of the centres are sparse, and their dimension is possibly much larger than the sample size. Our theoretical analysis focuses on the simple model $X_i = z_i\theta + \varepsilon_i$, $z_i \in \{-1,1\}$, $\varepsilon_i \sim \mathcal{N}(0, I_p)$ which has two clusters with centres θ and $-\theta$. We provide a finite sample analysis of a new sparse clustering algorithm based on sparse PCA and show that it achieves the minimax optimal misclustering rate in the regime $\|\theta\| \to \infty$, matching asymptotically the Bayes error. The results require the sparsity to grow slower than the square root of the sample size. Using a recent framework for computational lower bounds—the low-degree likelihood ratio—we give evidence that this condition is necessary for any polynomial-time clustering algorithm to succeed below the BBP threshold. This complements existing evidence based on reductions and statistical query lower bounds. Compared to these existing results, we cover a wider set of parameter regimes and give a more precise understanding of the runtime required and the misclustering error achievable. We also discuss extensions of our results to more than two clusters.

EO075 Room R20 RECENT ADVANCES IN STOCHASTIC NETWORK MODELS Chair: Marianna Pensky

E0309: **Improved clustering algorithms for the bipartite stochastic block model**
Presenter: **Suzanne Sigalla**, CREST, ENSAE, France
Co-authors: Alexandre Tsybakov, Mohamed Ndaoud

Consider a Bipartite Stochastic Block Model (BSBM) on vertex sets V1 and V2. We investigate sufficient conditions of exact and almost full recovery for clustering over V1 using polynomial-time algorithms, in the regime where the size of V1 is way smaller than the size of V2. We improve upon the known conditions of almost full recovery for spectral clustering algorithms in BSBM. Furthermore, we propose a new computationally simple and fast procedure achieving exact recovery under milder conditions than the state of the art. This procedure is a variant of Lloyd's iterations initialized with a well-chosen spectral algorithm leading to what we expect to be optimal conditions for exact recovery in this model. The latter fact is further supported by showing that a supervised oracle procedure requires similar conditions to achieve exact recovery. The key elements of the proof techniques are different from classical community detection tools on random graphs. Numerical studies confirm our theory, and show that the suggested algorithm is both high-speed and achieves similar performance as the supervised oracle. Finally, using the connection between planted satisfiability problems and the BSBM, we improve upon the sufficient number of clauses to completely recover the planted assignment.

E0324: **Extended stochastic block models**
Presenter: **Daniele Durante**, Bocconi University, Italy
Co-authors: Sirio Legramanti, Tommaso Rigon, David Dunson

Stochastic block models are widely used in network science due to their interpretable structure that allows inference on groups of nodes having common connectivity patterns. Although providing a well established model-based approach for community detection, such formulations are still the object of intense research to address the problem of inferring the unknown number of communities. This has motivated the development of several probabilistic mechanisms to characterize the node partition process, covering solutions with a fixed, random and infinite number of communities. We will provide a unified view of all these formulations within a single extended stochastic block model (ESBM), that relies on Gibbs-type processes and encompasses most existing representations as special cases. Connections with Bayesian nonparametric literature open new avenues that allow the inclusion of unexplored options to model the nodes partition process and to incorporate node attributes. Among these new alternatives, we focus on the Gnedin process as an example of a probabilistic mechanism with desirable properties and empirical performance. A collapsed Gibbs sampler that can be applied to the whole ESBM class is proposed, and refined methods for posterior inference are outlined. The performance of ESBM is assessed in simulations and an application to political networks.

E0737: **Mixed membership stochastic blockmodels for heterogeneous networks**
Presenter: **Yuguo Chen**, University of Illinois at Urbana-Champaign, United States

Heterogeneous networks are useful for modeling complex systems that consist of different types of objects. We formulate a heterogeneous version of the mixed membership stochastic blockmodel to accommodate heterogeneity in the data and the content dependent property of the pairwise relationship. We also apply a variational algorithm for posterior inference. The proposed procedure is shown to be consistent for community detection under mixed membership stochastic blockmodels for heterogeneous networks. We demonstrate the advantage of the proposed method in modeling overlapping communities and multiple memberships through simulation studies and applications to a real data set.

E0861: **A statistical interpretation of spectral embedding: The generalised random dot product graph**
Presenter: **Joshua Cape**, University of Pittsburgh, United States
Co-authors: Patrick Rubin-delanchy, Minh Tang, Carey Priebe

The purpose is to introduce the generalised random dot product graph model, a latent space network model that provides a unified setting in which to study spectral clustering methods, factorizations of matrices, and the geometry of point cloud configurations. We establish that, for both the normalised Laplacian and adjacency matrix, the vector representations of nodes obtained by spectral embedding provide strongly consistent latent position estimates with asymptotically Gaussian error. Direct methodological consequences follow from the observation that mixed membership and standard stochastic block models are special cases where the latent positions live respectively inside or on the vertices of a simplex. Hence, estimation via spectral embedding can be achieved by estimating the simplicial support or by fitting a Gaussian mixture model, respectively. We highlight applications of our results to the analysis of cybersecurity data and connectomics.

E0459: **Sparse popularity adjusted stochastic block model**
Presenter: **Marianna Pensky**, University of Central Florida, United States

The recently-introduced Popularity Adjusted Block model (PABM) is considered. To the best of our knowledge, the PABM is the only stochastic block model that allows treating the network sparsity as the structural sparsity that describes community patterns, rather than being an attribute of the network as a whole.

EO740 Room R21 BIOSTATISTICS IN CANCER RESEARCH Chair: Christiana Kartsonaki

E0942: **Exposure stratified case-cohort studies**
Presenter: **Martyn Plummer**, University of Warwick, United Kingdom

The China Kadoorie Biobank (CKB) is a prospective study of over 510,000 adults in 10 regions of China. As part of the CKB we investigated the bacterium H. pylori as a cause of several gastrointestinal diseases: gastric non-cardia cancer, gastric cardia cancer, oesophageal cancer, and gastric and duodenal ulcer. Due to the high cost of the serological tests for H. pylori, a sample of the CKB has been considered. A case-cohort sample

was chosen as an appropriate design when there are multiple disease outcomes of interest. We also wanted to optimize the selection of controls to account for the substantial variation in disease risk by age and sex, leading us to the exposure-stratified case-cohort design. The literature on exposure-stratified case-cohort designs contains relatively little advice on optimal design. We, therefore, chose our sub-cohort according to the well-established epidemiological principle of frequency matching. We also found unresolved questions in the analysis of these studies. Previous simulations have shown the superiority of Borgan III, the only unbiased estimating function for exposure stratified case-cohort studies. However, Borgan III is not fully efficient: the price of unbiasedness is that one control is discarded at random. We consider possible advances in the design and analysis of these studies.

E0888: **From genome-wide association studies to personalized risk prediction for breast cancer**
Presenter: **Kyriaki Michailidou**, The Cyprus Institute of Neurology and Genetics, Cyprus

Over the last decade, Genome-wide association studies (GWAS) have identified thousands of variants robustly associated with different complex traits. Large scale genotyping and meta-analyses of more than 200,000 breast cancer cases and controls, mainly through the Breast Cancer Association Consortium (BCAC), led to the identification of 172 loci that are associated with breast cancer susceptibility. The majority of these variants are common in the population (minor allele frequency of more than 5%) and contribute to a small fraction of the disease risk. These variants have shown to combine in a multiplicative way and can be summarized as a polygenic risk score (PRS) that can be used to categorize women into different risk categories and thus has the potential to be used in breast cancer prevention.

E1068: **Genomic analysis of breast cancers and biomarkers of therapeutic responsiveness**
Presenter: **Maggie Cheang**, The Institute of Cancer Research, United Kingdom

Various application of emerging technologies for biomarker research on FFPE materials will be discussed, in particular the integration of biomarker data from DNA, RNA, and proteomics-based assays for molecular classifications using breast cancer studies as the main examples. Various challenges and considerations taken in the study design and analytical methodologies when developing the PAM50 classifier will be shared. Several aspects to be considered will be discussed when planning a translational biomarker study with objectives to identify and develop genomics assays for sub-classification of tumours and to determine the sensitivity of each tumour type to therapeutic agents and targeted therapies.

E1010: **Applying machine learning methods to understand biological heterogeneity**
Presenter: **Yingtong Chen**, University of Oxford, United Kingdom
Co-authors: Maggie Cheang

Machine learning methods are applied to analyse genetics data for patients diagnosed with sarcoma. Firstly, unsupervised learning methods such as PCA, TSNE and k-means clustering are used to cluster the data into several subgroups. Then survival analysis is done based on these different groups to identify the difference of survival condition for patients in different groups. It is indeed found some subtypes perform better than other subtypes. Cox regression is also applied to analyse each gene to filter genes that affect survival conditions. Unsupervised learning methods to cluster patients based on the selected genes generate a clearer division among subgroups. Gene enrichment analysis software can also help find whether any cluster of genes is related to certain pathway/drug targets.

E1005: **DNA repair biology signatures to predict carboplatin (C) vs docetaxel (D) benefit in advanced TNBC**
Presenter: **Holly Tovey**, The Institute of Cancer Research, United Kingdom
Co-authors: Maggie Cheang

In the TNT Trial no improved response rate (RR) to C over D in aTNBC was observed, but it was in BRCA1/2 mutated patients (pts). We hypothesise tumours with other aberrant DNA damage response (DDR) characteristics have higher RR to C than D. We tested the predictive value of DDR related gene expression signatures (PARPi7, chromosomal instability CIN70, TP53 & DDIR) on 192 treatment-naive primary tumours (PT) by total RNA-sequencing. Paired PT and recurrent (REC) signature scores were compared. PT-REC pairs were available from 12 pts who received chemotherapy (CT) between PT & REC. CIN70 increased from PT to REC, DDIR (non-significantly) & PARPi7 decreased. 4/5 TP53 wildtype classified PT samples classified as mut in REC. DDIR predicted response to D in pts who received CT before trial entry. PARPi7 predicted response to D in CT naive pts. In CT naive pts, high CIN70 tumours suggested higher C RR as hypothesised. In conclusion, in this trial of aTNBC, DDIR high pts with prior CT had better RR to D than C. A possible explanation for this unexpected result is selective pressure of adjuvant DNA damaging CT and selection for relative taxane sensitivity in those who recur despite a high DDIR score. The hypothesised CIN70 treatment interaction was observed in CT naive pts. Our results suggest care is required in the application of signatures to initial diagnostic material when predicting response to DNA damaging agents at REC particularly in pts with prior CT.

EO514 Room R22 ADAPTIVE BAYESIAN METHODS FOR TIME AND SPATIAL SERIES ANALYSIS Chair: Scott Bruce

E0575: **Bayesian analysis of nonstationary periodic time series**
Presenter: **Beniamino Hadj-Amar**, Rice University, United States
Co-authors: Mark Fiecas, Barbel Finkenstadt, Francis Levi, Robert Huckstepp

Statistical methodology for identifying periodicities in cyclical phenomenon allows us to gain insight into the sources of variability that drive such a phenomenon. Non-stationary behavior seems to be the norm rather than the exception in physiological time series as time-varying periodicities, and other forms of rich dynamical patterns are commonly observed. We address these challenges and present two novel Bayesian methodologies for the automated analysis of these types of data. First, we propose to approximate the time series using a piecewise oscillatory model with unknown periodicities, where our goal is to estimate the change-points while simultaneously identifying the potentially changing periodicities in the data. Second, we present a non-parametric HMM where the states are defined through the spectral properties of a periodic regime. This approach further quantifies the probabilistic mechanism governing the transitions and recurrence of distinct periodic patterns. We show that the proposed methodologies are successfully applied in several applications that are relevant to e-Health and sleep research.

E0918: **Robust conditional spectral analysis of replicated time series**
Presenter: **Zeda Li**, City University of New York, United States

Traditional spectral analysis, which is based on the Fourier transform of the autocovariance, focuses on summarizing the cyclical behavior of a single time series. However, this type of analysis is subject to two major limitations: first, being covariance-based, it cannot accommodate heavy-tail dependence and infinite variance, and detect dynamics in time-irreversibility and kurtosis; second, focusing on a single time series, it is unable to analyze multiple time series and to quantify how their spectra are associated with other variables. We propose a new nonparametric approach to the spectral analysis of multiple time series and the associated covariates. The procedure is based on the copula spectral density kernel, which inherits the robustness properties of classical quantile regression and does not require any distributional assumptions, such as the existence of finite moments. Copula spectral density kernel of different pairs are modeled jointly as a matrix to allow flexible smoothing. Through a tensor-product spline model of Cholesky components of outcome-dependent copula spectral densities, the approach provides flexible nonparametric estimates of copula spectral density matrix as nonparametric functions of frequency and outcome while preserving geometric constraints.

E0973: **AdaptSPEC-X: Covariate dependent spectral modeling of multiple nonstationary time series**
Presenter: **Sally Cripps**, University of Sydney, Australia
Co-authors: Michael Bertolacci, Ori Rosen, Edward Cripps

A method for the joint analysis of a panel of possibly nonstationary time series is presented. The approach is Bayesian and uses a covariate-

dependent infinite mixture model to incorporate multiple time series, with mixture components parameterized by a time-varying mean and log spectrum. The mixture components are based on AdaptSPEC, a nonparametric model which adaptively divides the time series into an unknown but nite number of segments and estimate the local log spectra by smoothing splines. We extend AdaptSPEC to handle missing values, a common feature of time series which can cause difficulties for nonparametric spectral methods. A second extension is to allow for a time-varying mean. Covariates, assumed to be time-independent, are incorporated via the mixture weights using the logistic stick-breaking process. The resulting model can estimate time-varying means and spectra at both observed and unobserved covariate values, allowing for predictive inference. Estimation is performed by Markov chain Monte Carlo (MCMC) methods, combining data augmentation, reversible jump, and Riemann manifold Hamiltonian Monte Carlo techniques. We evaluate the methodology using simulated data and describe applications to Australian rainfall data and measles incidence in the US. Efficient software implementing the proposed method is available in the R package BayesSpec.

E0907: **A scalable partitioned approach to model massive non-stationary non-gaussian spatial data**
Presenter: **Seiyon Lee**, George Mason University, United States
Co-authors: Jaewoo Park

Nonstationary non-Gaussian spatial data are common in many disciplines, including the environmental sciences, ecology, epidemiology, and social sciences. Examples include count data on disease incidence and binary satellite data on cloud mask (cloud/no-cloud). Modeling such data sets as global stationary spatial processes can be unrealistic since they are collected over large heterogeneous domains (i.e. range of spatial dependence varies across subregions). Although several approaches have been developed for nonstationary spatial models, these have focused primarily on Gaussian data. In fact, fitting nonstationary models to large non-Gaussian data sets can be computationally prohibitive. To address these challenges, we propose a scalable algorithm for modeling such data that leverages the parallel computing in modern high performance computing (HPC) systems. We partition the spatial domain into disjoint subregions and fit locally non-stationary models using a carefully curated set of spatial basis functions. Then, we combine the local processes using a novel adaptive neighbor-based weighting scheme. Our approach scales well to massive datasets (hundreds of thousands), provides accurate predictions, and can be implemented in the nimble software environment. We demonstrate our method to simulated examples and two large real-world data sets pertaining to infectious diseases and remotely sensed images of cloud cover.

E0870: **Adaptive Bayesian covariate dependent spectral analysis of multiple time series**
Presenter: **Yakun Wang**, George Mason University, United States
Co-authors: Zeda Li, Scott Bruce

A flexible and adaptive method is proposed for estimating the association between multiple covariates and the power spectrum of multiple time series. The proposed approach uses a Bayesian "sum-of-trees" model to capture complex dependencies and interactions between covariates and the power spectrum. Local power spectra corresponding to terminal nodes within trees are estimated nonparametrically using Bayesian penalized linear splines. The tree structures in this model are considered to be random and fit using a Bayesian backfitting Markov chain Monte Carlo (MCMC) algorithm that sequentially considers modifications to trees via reversible jump MCMC techniques. By averaging over the posterior distribution of tree structures to estimate the covariate-dependent power spectrum, the proposed method can recover both smooth and abrupt changes in the power spectrum across covariates. Empirical performance is evaluated via simulations, which demonstrates the proposed method's ability to accurately recover complex nonlinear associations and interaction effects on the power spectrum.

EO279 Room R23 ADVENTURES IN BAYESIAN NONPARAMETRICS Chair: Ramses Mena

E0413: **Bayesian clustering of high-dimensional data**
Presenter: **Antonio Canale**, University of Padua, Italy
Co-authors: Noirrit Kiran Chandra, David Dunson

In many applications, it is of interest to cluster subjects based on very high-dimensional data. Although Bayesian discrete mixture models are often successful at model-based clustering, we demonstrate pitfalls in high-dimensional settings. The first key problem is a tendency for posterior sampling algorithms based on Markov chain Monte Carlo to produce a very large number of clusters that slowly decreases as sampling proceeds, indicating serious mixing problems. The second key problem is that the true posterior also has aberrant behaviour but potentially in the opposite direction. In particular, we show that, for diverging dimension and fixed sample size, the true posterior either assigns each observation to a different cluster or all observations to the same cluster, depending on the kernels and prior specification. We propose a general strategy for solving these problems by basing clustering on a discrete mixture model for a low-dimensional latent variable. We refer to this class of methods as LAtent Mixtures for Bayesian (Lamb) clustering. Theoretical support is provided, and we illustrate substantial gains relative to clustering on the observed data level in simulation studies. The methods are motivated by an application to clustering of single cell RNAseq data, with the clusters corresponding to different cell types.

E0538: **The semi-hierarchical Dirichlet process and its application to clustering homogeneous distributions**
Presenter: **Alessandra Guglielmi**, Politecnico di Milano, Italy
Co-authors: Mario Beraha, Fernando Quintana

Assessing homogeneity of distributions is an old problem that has received considerable attention, especially in the nonparametric Bayesian literature. To this effect, we propose the semi-hierarchical Dirichlet process, a novel hierarchical prior that extends a previous hierarchical Dirichlet process and that avoids the degeneracy issues of nested processes recently described. We go beyond the simple yes/no answer to the homogeneity question and embed the proposed prior in a random partition model; this procedure allows us to give a more comprehensive response to the above question and in fact find groups of populations that are internally homogeneous when such populations (two or more) are considered. Simulation studies and an application to educational data are also discussed.

E0550: **Bayesian nonparametric hypothesis testing procedures**
Presenter: **Luis Gutierrez**, Pontificia Universidad Catolica de Chile, Chile

Scientific knowledge is firmly based on the use of statistical hypothesis testing procedures. A scientific hypothesis can be established by performing one or many statistical tests based on the evidence provided by the data. Given the importance of hypothesis testing in science, these procedures are an essential part of statistics. The literature on hypothesis testing is vast and covers a wide range of practical problems. However, most of the methods are based on restrictive parametric assumptions. We will discuss Bayesian nonparametric approaches to construct hypothesis tests in different contexts. Our proposal resorts to the literature of model selection to define Bayesian tests for multiple samples, paired-samples, and longitudinal data analysis. Applications with real-life datasets and illustrations with simulated data will be discussed.

E0640: **Stick-breaking processes with exchangeable length variables**
Presenter: **Maria Fernanda Gil-Leyva Villa**, IIMAS,UNAM, Mexico
Co-authors: Ramses Mena

The stick-breaking construction is a well-known method to define distributions on the infinite-dimensional simplex as well as Bayesian nonparametric priors. Due to the mathematical hurdles to overcome, most efforts of studying this class with some generality, have concentrated in the case where the underlying length variables are independent. The focus is on the general class of stick-breaking processes with exchangeable length variables. These generalize well-known Bayesian non-parametric priors, such as Dirichlet and Geometric processes, in an unexplored direction.

We give conditions to assure the respective species sampling process is discrete almost surely and the corresponding prior has full support. For a rich sub-class, we study the ordering of the stick-breaking weights and derive and an MCMC algorithm for density estimation purposes.

E0692: **Nonparametric Bayesian modelling of longitudinally integrated covariance functions on the sphere**
Presenter: **Bernardo Nipoti**, University of Milan Bicocca, Italy
Co-authors: Pier Giovanni Bissiri, Galatia Cleanthous, Xavier Emery, Emilio Porcu

Taking into account axial symmetry in the covariance function of a Gaussian random field is essential when the purpose is modelling data defined over a large portion of the sphere representing our planet. Axially symmetric covariance functions admit a convoluted spectral representation which makes modelling and inference difficult. This motivates the interest in devising alternative strategies to attain axial symmetry, an appealing option being longitudinal integration of isotropic processes on the sphere. We provide a comprehensive theoretical framework to model longitudinal integration on spheres through a nonparametric Bayesian approach. Longitudinally integrated covariances are treated as random objects, where the randomness is implied by the randomised spectrum associated with the covariance function. We then define and implement a Bayesian nonparametric model for the analysis of data defined on the sphere. We investigate its properties and assess its performance through the analysis of both simulated data and a data set on mean daily air temperatures extracted from the NCEP/NCAR Reanalysis 1 data set.

EO546 Room R24 CHALLENGES AND RECENT ADVANCES IN OPTIMAL DESIGN Chair: Saumen Mandal

E1008: **Design and analysis of adaptive clinical trials**
Presenter: **Christopher Jennison**, University of Bath, United Kingdom

Clinical trials are conducted at various stages of the drug development process. The aims at each stage are different: Phase I trials are designed to find the maximum tolerated dose, possibly studying an efficacy response as well; Phase II trials seek the most effective dose and evidence that, at this dose, the new treatment will be superior to the control; at the confirmatory stage, Phase III trials are conducted to demonstrate the superiority of the new treatment with respect to the primary endpoint. In order to create an efficient design for a given trial, the goal of that trial needs to be specified and this requires a clear view of the role of the trial in the wider process. We shall present examples of such model-based decision making and the development of efficient experimental designs for: dose escalation in a Phase I trial; the transition from Phase II to Phase III trials; adaptive enrichment in a Phase III trial.

E0475: **Design selection and analysis for two-level supersaturated designs**
Presenter: **Rakhi Singh**, UNC Greensboro, United States
Co-authors: John Stufken

An extensive literature is available on design selection criteria and analysis techniques for 2-level supersaturated designs. The most notable design selection criteria are the popular $E(s^2)$-criterion, $UE(s^2)$-criterion, and Bayes D-optimality criterion, while the most notable analysis technique is the Gauss-Dantzig Selector. It has been observed that while the Gauss-Dantzig Selector is the superior analysis technique, differences in screening performance of different designs are not captured by any of the common design selection criteria. We will consider new design selection criteria inspired by the Gauss-Dantzig Selector. We will establish that designs that are better under these criteria also tend to perform better as screening designs. In addition, most supersaturated designs are studied under the main effects model because the number of runs is small to study even the main effects, let alone be studying the main effects as well as the two-factor interactions. However, in practice, there is no guarantee that the presence of two-factor interactions does not influence the response. We will also consider a random forest-inspired analysis technique that we have developed to circumvent this problem. We will see that this technique outperforms the existing analysis techniques.

E0703: **Blocked foldover designs with column permutations**
Presenter: **Po Yang**, University of Manitoba, Canada
Co-authors: William Li

Follow-up experimentation is often necessary for the successful use of fractional factorial designs. Foldover is one of the techniques used in follow-up experiments. Since follow-up experiments are usually conducted in different stages, the blocking effect is often considered. We consider foldover designs with column permutation when a block factor is included. It is shown that a pair of a permutation plan and a fold-over plan has generalized minimum aberration for the blocked combined foldover design if and only if it has generalized minimum aberration for the unblocked combined foldover design.

E1090: **Designing to estimate parameters independently of each other**
Presenter: **Saumen Mandal**, University of Manitoba, Canada
Co-authors: Ben Torsney, Mohammad Chowdhury

An optimal design problem is considered where the criterion function is neither convex nor concave. Determining optimality conditions for such criteria is quite challenging. Motivated by this fact, we first establish two sets of optimality conditions for a non-concave criterion using Lagrangian theory and directional derivatives. We then apply these conditions to construct optimal designs for some regression models in which it is desired to estimate certain parameters as independently of each other as possible. We minimize absolute covariances among the least-squares estimators of the parameters in a linear model, thereby rendering the parameter estimators approximately uncorrelated with each other. We then consider the problem of obtaining more than two parameter estimators with more than one zero correlation among them. We achieve this goal by creating a compound criterion and then solving the problem by employing a simultaneous optimization technique. More specifically, we transform the problem to an optimization problem in which we maximize a number of functions of the design weights simultaneously. The methodologies are formulated for a general regression model and are explored through some examples, including one practical problem arising in chemistry.

E0609: **Exact permutation/randomization tests algorithms**
Presenter: **Subir Ghosh**, University of California, United States

R. A. Fisher described the exact permutation and randomization tests for comparative experiments without assuming normality or any particular probability distribution. While having this as an attractive feature, the computational challenge was a disadvantage at that time but not now with modern computers. A permutation/randomization data algorithm is introduced to generate the permutation/randomization distributions under the null hypotheses for calculating the P-values. The properties of permutation/randomization data matrices developed by algorithms following the proposed mathematical processes are derived. Two illustrative examples demonstrate the usefulness of the proposed computational methods.

EO744 Room R25 RECENT DEVELOPMENT OF SUFFICIENT MULTIVARIATE METHODS Chair: Chenlu Ke

E0262: **On sufficient graphical models**
Presenter: **Kyongwon Kim**, Wake Forest University, United States

A sufficient graphical model is introduced by applying the recently developed nonlinear sufficient dimension reduction techniques to the evaluation of 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, and its performance remains excellent in the high-dimensional setting.

E0526: **Parsimonious multivariate spatial regression**
Presenter: **Hossein Moradi Rekabdarkolaee**, South Dakota State University, United States

Dimension reduction provides a useful tool for analyzing high dimensional data. The recently developed Envelope method is a parsimonious version of the classical multivariate regression model by identifying a minimal reducing subspace of the responses. We introduce an extension of the envelope, called spatial envelope method, for dimension reduction in the presence of dependencies across space. We studied the effectiveness o this approach through a simulation study and data analysis.

E0762: **On sufficient variable screening using log odds ratio filter**
Presenter: **Wenbo Wu**, University of Texas at San Antonio, United States

For ultrahigh-dimensional data, variable screening is an important step to reduce the scale of the problem, hence, to improve the estimation accuracy and efficiency. We propose a new dependence measure which is called the log odds ratio statistic to be used under the sufficient variable screening framework. The sufficient variable screening approach ensures the sufficiency of the selected input features in modeling the regression function and is an enhancement of existing marginal screening methods. 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 advantages 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. Extensive simulations and a real data analysis are provided to demonstrate the usefulness of the proposed log odds ratio filter and the sufficient variable screening procedure.

E0969: **A sufficient dimension reduction method via expectation of conditional difference**
Presenter: **Qingcong Yuan**, Miami University, United States

An approach is introduced to sufficient dimension reduction problems using an expectation of conditional difference measure. The proposed method requires very mild conditions on the predictors, estimates the central subspace effectively and is especially useful when the response is categorical. It keeps the model-free advantage without estimating link function. Under regularity conditions, root-n consistency and asymptotic normality are established. The proposed method is very competitive and robust comparing to existing dimension reduction methods through simulations results.

E1169: **Nonparametric tests for multivariate data with missing values**
Presenter: **Yue Cui**, Missouri State University, United States

Quality of Life (QOL) outcomes are important in the management of chronic illnesses. In studies of efficacies of treatments or intervention modalities, QOL scales-multi-dimensional constructs are routinely used as primary endpoints. The standard data analysis strategy computes composite (average) overall and domain scores, and conducts a mixed-model analysis for evaluating efficacy or monitoring medical conditions as if these scores were in continuous metric scale. However, assumptions of parametric models like continuity and homoscedasticity can be violated in many cases. Furthermore, it is even more challenging when there are missing values on some of the variables. We will introduce a purely nonparametric approach in the sense that meaningful and, yet, nonparametric effect size measures are developed. We propose an estimator for the effect size and develop the asymptotic properties. The methods are shown to be, particularly effective in the presence of some form of clustering and/or missing values. Inferential procedures are derived from the asymptotic theory. The Asthma Randomized Trial of Indoor Wood Smoke data will be used to illustrate the applications of the proposed methods. The data was collected from a three-arm randomized trial which evaluated interventions targeting biomass smoke particulate matter from older model residential wood stoves in homes that have kids with asthma.

EC800 Room R02 CONTRIBUTIONS IN METHODOLOGICAL STATISTICS Chair: Fatemeh Ghaderinezhad

E0511: **Censored Poisson regression with missing censoring information**
Presenter: **Jean-Francois Dupuy**, INSA de Rennes, France
Co-authors: Bilel Bousselmi, Abderrazek Karoui

Estimation in the Poisson regression model is considered when the count response is randomly right-censored and the censoring indicator can be missing at random. We investigate several estimation methods, such as multiple imputation and augmented-inverse-probability-weighted estimation. We derive the asymptotic properties of the resulting estimators (consistency, asymptotic normality, consistent variance estimation). A simulation study is conducted to evaluate and compare the proposed estimates.

E0644: **Assessing input variable activity for Bayesian regression trees**
Presenter: **Akira Horiguchi**, The Ohio State University, United States

Bayesian Additive Regression Trees (BART) are non-parametric models that can capture complex exogenous variable effects. In any regression problem, it is often of interest to learn which variables are most active. Variable activity in BART is usually measured by counting the number of times a tree splits for each variable. Such one-way counts have the advantage of fast computations. Despite their convenience, one-way counts have several issues. They are statistically unjustified, cannot distinguish between main effects and interaction effects, and become inflated when measuring interaction effects. An alternative method well-established in the literature is Sobol´ indices, a variance-based global sensitivity analysis technique. However, these indices often require Monte Carlo integration, which can be computationally expensive. Analytic expressions for Sobol´ indices for BART posterior samples are provided. These expressions are easy to interpret and are computationally feasible. Furthermore, we will show a fascinating connection between first-order (main-effects) Sobol´ indices and one-way counts. We also introduce a novel ranking method and use this to demonstrate that the proposed indices preserve the Sobol´-based rank order of variable importance. Finally, we compare these methods using analytic test functions and the En-ROADS climate impacts simulator.

E0826: **Semiparametric prediction intervals in parametric models with non-normal additive error terms**
Presenter: **Gerhard Fechteler**, Universitat Konstanz, Germany

The asymptotic distribution of mean predictions in parametric regression models with additive error term structure is usually well known and often normal. The construction of confidence intervals based on the asymptotic distribution is straight forward. To account for the uncertainty resulting from the error terms, prediction intervals are often more meaningful in applied work. Prediction intervals are commonly constructed for normally distributed error terms in the literature. We propose a simple framework for constructing prediction intervals for non-normal error term distributions. We show that the interval is based on a distribution resulting from the convolution of the distributions of the mean prediction and the error term. The estimation strategy is based on a kernel density estimation of the error term distribution. The implementation is straight forward and applicable to all regression models with known (asymptotic) parameter distribution. We demonstrate the usefulness of the framework via an application to the prediction of house prices.

E0862: **Inferences for the correct classification fractions of a continuous biomarker in trichotomous settings**
Presenter: **Peng Shi**, University of Kansas Medical Center, United States
Co-authors: Leonidas Bantis

Hepatocellular carcinoma (HCC) is the most common primary cancer of the liver. As such, there is a strong clinical interest in finding new biomarkers for its early detection. When the disease status is trichotomous, the ROC surface is an appropriate tool for assessing the discriminatory ability of a marker. A popular approach for computing cutoffs for decision making is the Youden index and its recent 3-class generalization.

However, this method treats the data in a pairwise fashion and is unable to accommodate biomarker scores from all three groups simultaneously. This may result in inappropriate cutoffs that are of no clinical interest. Methods are proposed for such inferences where the cutoffs are based on the minimized Euclidean distance of the ROC surface from the perfection corner. An inferential framework, both parametric and non-parametric, are provided for the derivation of marginal confidence intervals (CIs) and joint confidence spaces (CSs) for the optimized true class rates. Our approaches were evaluated through extensive simulations and finally illustrated using a real data set that refers to HCC patients.

E0948: **Estimation and construction of CIs for the cut points of cont. biomarkers under the Euclidean distance in 3D settings**
Presenter: **Brian Mosier**, University of Kansas Medical Center, United States
Co-authors: Leonidas Bantis

Pancreatic ductal adenocarcinoma (PDAC) is an aggressive type of cancer with a 5-year survival rate of less than 5%. As in many other diseases, its diagnosis might involve progressive stages. It is common that in biomarker studies referring to PDAC, recruitment involves three groups: healthy individuals, patients that suffer from chronic pancreatitis and PDAC patients. Early detection and accurate classification of the state of the disease are crucial for patients' successful treatment. ROC analysis is the most popular way to evaluate the performance of a biomarker, and the Youden index is commonly employed for cutoff derivation. The so-called generalized Youden index has a drawback in the three-class case of not accommodating the full data set when estimating the optimal cutoffs. We explore the use of the Euclidean distance of the ROC to the perfection corner for the derivation of cutoffs in trichotomous settings. We construct an inferential framework that involves both parametric and non-parametric techniques. The proposed methods can accommodate the full information of a given data set and thus provide more accurate estimates in terms of the decision-making cutoffs compared to a Youden-based strategy. We evaluate our approaches through extensive simulations and illustrate them on a PDAC biomarker study.

CI025 Room R04 ECONOMETRIC CHALLENGES CAUSED BY PANDEMIC Chair: Andrew Butters

C0162: **How to estimate a VAR after March 2020**
Presenter: **Giorgio Primiceri**, Northwestern University, United States

The aim is to illustrate how to handle a sequence of extreme observations–such as those recorded during the COVID-19 pandemic–when estimating a Vector Autoregression, which is the most popular time-series model in macroeconomics. The results show that the ad-hoc strategy of dropping these observations may be acceptable for the purpose of parameter estimation. However, disregarding these recent data is inappropriate for forecasting the future evolution of the economy, because it vastly underestimates uncertainty.

C1171: **Macroeconomic forecasting during disaster recovery**
Presenter: **Jeffrey Campbell**, University of Notre Dame, United States

An empirical model of ongoing disaster recovery is provided to augment a preexisting Box-Jenkins forecasting model. The model sums the original series with possibly random disaster shocks, each of which follows a first-order autoregression. Application of the Kalman filter yields forecasts of the disaster recovery's expected duration and estimates of the counterfactual time series without the disaster. Applications to Japanese IP following the Tohoku Earthquake, Israeli IP following the Yom Kippur War and the U.S. and European macroeconomic time series since March 2020 illustrate the model's usefulness.

C0164: **This time is different: Disentangling the channels of the Covid-19 recession**
Presenter: **Andrew Butters**, Indiana University, United States
Co-authors: Scott Brave

The aim is to examine the macroeconomic dynamics of the recession induced by private and public sector responses to the Covid-19 pandemic in the U.S. Using a large dataset of macroeconomic indicators maintained by the St. Louis Fed (FRED-MD), we show that the unprecedented declines in economic activity in the spring and subsequent rebound in the summer months of 2020 pose a considerable challenge to the use of factor models to consistently capture the macroeconomic dynamics of both the Covid-19 and past U.S. recessions. Unlike what was found for the 2007-09 (Great) recession previously, the dynamics of the Covid-19 recession were very different than prior recessions in the U.S. This result is manifested in a structural break in the factor loadings of single and multifactor representations of the FRED-MD dataset. We describe a simple reduced-form correction that can be made to standard factor models estimated by PCA which isolates the common business cycle variation in economic activity from the idiosyncratic volatility specific to the Covid experience, producing consistent factor loadings and factors throughout the period from the early 1960s through the present.

CO113 Room R03 REGIME CHANGE I: BUSINESS CYCLES AND REGIME CHANGE Chair: Ibrahim Tahri

C0444: **Endogenous and exogenous volatility in the foreign exchange market**
Presenter: **Leonardo Bargigli**, Universita di Firenze, Italy

Two sources of heteroskedasticity in high-frequency financial data are identified. The first source is the endogenous changing participation of heterogeneous speculators to the market, coupled with the time-varying behaviour of the market maker. The second source is the exogenous flow of market-relevant information. We model the first one using a Markov switching (MS) SVAR process and the second one utilizing a GARCH process for the MS-SVAR structural errors. Using transaction data of the EUR-USD market in 2016, we detect three regimes characterized by different levels of endogenous volatility. The impact of structural shocks on the market depends on both sources, but the exogenous information is channelled to the market mostly through price. This suggests that the market maker is better informed than the speculators, who act as momentum traders. The latter can profit from trade because, unlike noise traders, they respond immediately to price shocks.

C0455: **Understanding business cycles in the perspectives of stabilizing and destabilizing mechanism**
Presenter: **Gang Gong**, Yunnan University of Finance and Economics, China

The so-called stabilization mechanism is a type of economic activity embedded in the economy that can stabilize it when it deviates from equilibrium (or a steady-state). A typical example is price adjustment. While price adjustment can be seen as a stabilization mechanism, are there mechanisms that destabilize an economy? The aim is to demonstrate that investment adjustment is a destabilizing mechanism. Given the existence of a destabilization mechanism, economic fluctuations (or business cycles) can thus be understood as the interaction between these two mechanisms. We build a macro-dynamic model with investment and price as the core macroeconomic variables. The analysis shows that due to price stickiness, the price adjustment mechanism may not be enough to stabilize the economy. In this case, a government stabilization policy is necessary for further stabilization. We also provide an empirical analysis to test our theoretical results.

C0449: **On the behavior of Okun's law across business cycles**
Presenter: **Luiggi Donayre**, University of Minnesota - Duluth, United States

Using threshold regression analysis, the aim is to study how the relationship between unemployment and output, known as Okun's law, varies across business cycles measures. At the U.S. aggregate level, the results indicate the presence of strong asymmetries in the relationship characterized by three different regimes, where the sensitivity of unemployment to output decreases with economic activity for all measures of business cycles considered. For the preferred threshold variable, Okun's law is strong when the unemployment gap grows faster than 1.07 percentage points above the natural rate of unemployment, coinciding with periods of deep recessions. The sensitivity is smaller in absolute value when it grows between -0.70 and 1.07 percentage points during mild recessions, and weakens even further when the unemployment gap falls below -0.70 percentage points

during periods of expansion, revealing a flattening, but also a shift, of Okun's law. The results are robust to the measure of the business cycle, the specification of Okun's law, the speed of output growth, the frequency of the data and the identification of the gaps. The analysis also finds support for the nonlinear nature of Okun's law at the state and international levels.

C0645: **The green transition: Directed technical change towards decarbonization**
Presenter: **Oriol Valles Codina**, UCL Institute for Innovation and Public Purpose, New School for Social Research, United States

The Flaschel-Semmler dynamical model of multi-sectorial growth is employed to study the substitution of carbon sectors by green sectors, that is, the Green Transition, under directed technical change. The Flaschel-Semmler model of linear production is based on the dynamic cross-dual linear adjustment between prices and quantities in the form of a law of excess demand and law of excess profitability, which produces a complex pattern of oscillations around their equilibrium values. The linear adjustment coefficients of the model are empirically calibrated for six countries using EU KLEMS and WIOD data. The paper concludes by evaluating analytically and computationally the tax rates that allow decarbonization to meet the targets of the UN Intergovernmental Panel on Climate Change. Directed technical change is enforced by a revenue-neutral, pro-active fiscal policy of a tax-subsidy form, which affects to greatly accelerate the phase-out of the carbon sector, in particular at its earliest stages. Without fiscal policy, no economy can reach the IPCC targets.

C0885: **A coordination game of regime change and democratization: Information manipulation and intermediate regimes**
Presenter: **Tariq Basir**, South Asian University, India
Co-authors: Soumya Datta

The purpose is to incorporate the collective action problem into a democratization framework which does not account for the collective action problem on the citizens' side in organizing themselves to impose a revolutionary threat to the elites/regime. We try to account for the collective action problem in this context, within a framework of global games of regime change. The preliminary analyzes suggest a unique equilibrium when we do not account for the engagement of the regime in information manipulation regarding the cost of revolution, i.e. signaling game; While it suggests the existence of multiple equilibria when we allow for information manipulation. We show that information manipulation becomes possible for the intermediate regime types only, and these regimes might be able to reduce the attack size and frequency of regime changes, by signaling a higher cost of revolution; which might not have been possible without the possibility of information manipulation (policy-intervention). Furthermore, the manipulation becomes possible when the opportunity cost of revolution is not very low (not in a severe economic downturn), implying when citizens are not able to solve the coordination problem among themselves easily. Similarly, the multiplicity result in our paper could be linked to the insecure autocracy/intermediate regimes literature for a richer comparative statics analysis.

CO720 Room R07 TOPICS OF MACHINE LEARNING AND ECONOMETRICS IN MONETARY POLICY Chair: Massimo Guidolin

C0211: **Federal reserve communications sentiment's impact on target rate discovery**
Presenter: **Juan Carlos Arismendi-Zambrano**, Maynooth University, Ireland
Co-authors: Massimo Guidolin, Alessia Paccagnini

The aim is to construct a communication risk profile of the US Federal Reserve Chair by measuring the sentiment of their public statements during their tenure. Statements' sentiment is calculated by a Naive Bayes text categorization method. Communications' sentiment impact on the interest rates price discovery process by the market after the FOMC meeting is analyzed. The results show that there is a significant difference in the communications' sentiment that plays a role in diminishing the volatility of Federal Reserve announcements and that can be effectively used as a tool for a monetary policy shocks.

C0213: **Estimating the effects of monetary policy: An automated narrative approach**
Presenter: **Miguel Acosta**, Columbia University, United States

The aim is to investigate whether the macroeconomic effects of monetary policy can be separately identified from the effects of the information revealed by monetary policy actions and announcements. This task is complicated by the fact that both policies are emitted at the same time; thus, current estimates of monetary policy shocks are generally unable to distinguish between the two. An empirical strategy is presented for distinguishing between these effects. An information shock is identified via an "automated narrative approach". Specifically, Fed watchers describe what information was revealed in each policy announcement. By applying machine learning and natural language processing techniques, the surprise component of the Fed's discussion about different topics (e.g., inflation or output) in its post-meeting announcements can be measured. These measures can be used as detailed proxies for the information content of monetary policy in order to study and estimate the effects of monetary policy and communications policy on financial market and macroeconomic outcomes.

C1112: **Narrative fragmentation and the business cycle**
Presenter: **Isaiah Hull**, Sveriges Riksbank, Sweden
Co-authors: Xin Zhang, Christoph Bertsch

According to previous work, economic and financial narratives often emerge as a consequence of their virality, rather than their veracity, and constitute an important, but understudied driver of aggregate fluctuations. Using a unique dataset of newspaper articles over the 1950-2019 period and state-of-the-art methods from natural language processing, we characterize the properties of business cycle narratives. The main finding is that narratives tend to consolidate around a dominant explanation during expansions and fragment into competing explanations during contractions. We also show that the existence of past reference events is strongly associated with increased narrative consolidation.

C0908: **Tweeting on monetary policy and market reactions**
Presenter: **Davide Romelli**, Trinity College Dublin, Ireland
Co-authors: Donato Masciandaro, Gaia Rubera

How does central bank communication affect financial markets? It is shown that monetary policy announcements of three major central banks, i.e. the Federal Reserve, the European Central Bank and the Bank of England, trigger a significant non-expert monetary policy discussion on Twitter. Using machine learning techniques, we classify Tweets related to monetary policy around the announcement date and build a metric of the similarity between the policy announcement and Twitter traffic before and after the announcement. We interpret large changes in the similarity of Tweets as a proxy for monetary policy surprise and show that market volatility spikes after the announcement whenever changes in similarity are high. These findings suggest that social media discussions on central bank communication are aligned with bond and stock market reactions.

C1130: **How much information monetary policy committees disclose: Evidence from the FOMCs minutes and transcripts**
Presenter: **Marianna Blix Grimaldi**, Swedish National Debt Office, Sweden
Co-authors: Isaiah Hull, Mikael Apel

The purpose of central bank minutes is to give an account of monetary policy meeting discussions to outside observers, thereby enabling them to draw informed conclusions about future policy. However, minutes are a shortened and edited representation of a broader discussion by necessity. Consequently, they may omit information that is predictive of future policy decisions. To investigate this, we compare the predictive content of the FOMCs minutes and transcripts, focusing on three dimensions which are likely to be excluded from the minutes: 1) the committee's degree of hawkishness, 2) the chairperson's degree of hawkishness, and 3) the level of agreement between committee members. We measure committee and chairperson hawkishness with a novel dictionary that is constructed using the FOMCs minutes and transcripts. The agreement is measured using a novel technique that we import from the machine learning literature. We show that transcripts contain information that is not included in

the minutes and is not contained in macroeconomic and financial variables. We also add evidence that the FOMC attempts to build an internal consensus before tightening monetary policy.

CO299 Room R08 LOCAL PROJECTIONS AND APPLICATIONS Chair: Alessia Paccagnini

C0404: **Assessing macroeconomic tail risk**
Presenter: **Christian Matthes**, Indiana University, United States
Co-authors: Francesca Loria, Donghai Zhang

GDP and Industrial Production in the US feature substantial tail risk. While this fact is well documented, it is not clear what drives this asymmetry - is there a common propagation mechanism or is it one specific shock only that drives this skewness? We provide evidence for the first explanation by (i)showing that the 10th percentile of the GDP and IP distributions responds drastically more to various supply and demand shocks than the median or the 90th percentile, and(ii) showing that two data-generating processes that feature a common propagation mechanism can match these patterns.

C0435: **Time-varying local projections-IV**
Presenter: **Germano Ruisi**, Central Bank of Malta, Malta

In recent years local projections have become a more and more popular methodology for the estimation of impulse responses. Besides being relatively easy to implement, the main strength of this approach, relative to the traditional VAR one, is that there is no need to impose any specific assumption on the dynamics of the data. In addition, the recent applied literature has also developed several instruments aiming at identifying several macroeconomic shocks of interest. Local projections-IV in a time-varying framework are modeled, and a Gibbs sampler routine is provided to estimate them. A simulation study shows how the performance of the algorithm is satisfactory while the usefulness of the model developed here is shown through an application to fiscal policy shocks.

C0694: **Dealing with the statistical representation of DSGE models**
Presenter: **Alessia Paccagnini**, University College Dublin, Ireland

Dynamic Stochastic General Equilibrium (DSGE) models are the main tool used in Academia and in Central Banks to evaluate the business cycle for policy and forecasting analyses. Despite the recent advances in improving the fit of DSGE models to the data, the misspecification issue remains. We deal with a specific aspect of the misspecification: the statistical representation of DSGE models. In particular, we discuss the case of DSGE models with a Vector Autoregressive Moving Average (VARMA) representation as a Data Generation Process. Considering several DSGE models to generate artificial pseudo-data, we compare results identifying shocks with VAR and Local Projection. We focus on the estimation and truncation errors induced by relying on a misspecified statistical representation.

C0695: **Local projection inference is simpler and more robust than you think**
Presenter: **Mikkel Plagborg-Moller**, Princeton University, United States
Co-authors: Jose Luis Montiel Olea

Applied macroeconomists often compute confidence intervals for impulse responses using local projections, i.e., direct linear regressions of future outcomes on current covariates. The aim is to prove that local projection inference robustly handles two issues that commonly arise in applications: highly persistent data and the estimation of impulse responses at long horizons. We consider local projections that control for lags of the variables in the regression. We show that lag-augmented local projections with normal critical values are asymptotically valid uniformly over (i) both stationary and non-stationary data, and also over (ii) a wide range of response horizons. Moreover, lag augmentation obviates the need to correct standard errors for serial correlation in the regression residuals. Hence, local projection inference is arguably both simpler than previously thought and more robust than standard autoregressive inference, whose validity is known to depend sensitively on the persistence of the data and on the length of the horizon.

C0865: **Bias in local projections**
Presenter: **Ed Herbst**, Federal Reserve Board, United States
Co-authors: Ben Johannsen

Local projections (LPs) are a popular tool in applied macroeconomic research. We survey the related literature and find that LPs are often used with very small samples in the time dimension. With small sample sizes, given the high degree of persistence inmost macroeconomic data, impulse responses estimated by LPs can be severely biased. This is true even if the right-hand-side variable in the LP is iid, or if the data set includes a large cross-section (i.e., panel data). We derive a simple expression for elucidating the source of the bias. Our expression highlights the interdependence between coefficients of LPs at different horizons. As a byproduct, we propose a way to bias-correct LPs.Using U.S. macroeconomic data and identified monetary policy shocks, we demonstrate that the bias correction can be large.

CC816 Room R06 CONTRIBUTIONS IN ECONOMETRIC MODELLING Chair: Maria Grith

C0229: **Technical efficiency and inefficiency: Reassurance of standard SFA models and a misspecification problem**
Presenter: **Anatoly Peresetsky**, National Research University Higher School of Economics, Russia
Co-authors: Subal Kumbhakar, Evgenii Shchetinin, Alexey Zaytsev

The purpose is to formally prove that if inefficiency (u) is modelled through the variance of u which is a function of z, then marginal effects of z on technical inefficiency (TI) and technical efficiency (TE) have opposite signs. This is true in the typical setup with normally distributed random error v and exponentially or half-normally distributed u for both conditional and unconditional TI and TE. We also provide an example to show that signs of the marginal effects of z on TI and TE may coincide for some ranges of z. If the real data comes from a bimodal distribution of u, and we estimate model with an exponential or half-normal distribution for u, the estimated efficiency and the marginal effect of z on TE would be wrong. Moreover, the rank correlations between the true and the estimated values of TE could be small and even negative for some subsamples of data. This result is a warning that the interpretation of the results of applying standard models to real data should take into account this possible problem. The results are demonstrated by simulations.

C1096: **Consumer theory with non-parametric taste uncertainty and individual heterogeneity**
Presenter: **Christopher Dobronyi**, University of Toronto, Canada
Co-authors: Christian Gourieroux

Two new classes of non-parametric random utility models for demand systems are introduced. In each class, individual-level heterogeneity is characterized by a distribution G over taste parameters, and heterogeneity across consumers is introduced by means of a distribution F over the distributions G. Demand is non-separable and heterogeneity is infinite-dimensional. Each class allows for corner solutions. We present two distinct frameworks for model estimation: (i) a Bayesian framework in which F is known, and (ii) a hyperparametric framework in which F is a member of a parametric family. We use a panel of scanner data to illustrate our methods in an application to the consumption of alcohol.

C0958: **Nonlinear impulse response function for dichotomous models**
Presenter: **Quentin Lajaunie**, Universita Paris Dauphine, France

A generalized impulse response function (GI) for dichotomous models is proposed. Building on previous work, we develop the exact form of the response functions for each specification of their binary model. Using a block-bootstrap method, we compute robust confidence intervals for these

response functions. We illustrate the usefulness of this analytical result for static and dynamic dichotomous models of U.S. recessions. According to the different specifications, we empirically find that the persistence of an impact of an exogenous shock to the U.S. economy is between one to five quarters.

C0535: **A new proposal for the construction of a multi-period/multilateral price index**

Presenter: **Consuelo Nava**, University of Aosta Valley, Italy

Co-authors: Maria Grazia Zoia

A price index providing a novel and effective solution both in a multi-period and a multilateral framework is devised within the stochastic framework. The derivation of the index, denoted MPL index, is obtained as a solution to an optimization problem which requires values and quantities of the commodities, not prices. Depending on the choice of the objective function to optimize, some of the most popular price indexes, namely Laspeyeres, Paasche, Marshall-Edgeworth and Walsh, arise as a special case. The MPL reference basket is the union of the intersections of the baskets of all periods/countries in pairs. As such, it provides broader coverage than usual indexes. Index closed-form expressions and updating formulas are provided and its properties investigated. Last, applications of the MPL index with both real and simulated data provide evidence of its good performance. In particular, a comparison between the MPL and the country/time-product dummy index highlights how the former proves to be more efficient with respect to the latter.

Saturday 19.12.2020 17:40 - 18:55 Parallel Session F – CFE-CMStatistics

EO690 Room R11 FUNCTIONAL DATA DEFINED OVER ARBITRARILY SHAPED DOMAINS Chair: Michelle Carey

E0939: **Non-parametric regression for networks**
Presenter: **Katie Severn**, University of Nottingham, United Kingdom
Co-authors: Ian Dryden, Simon Preston

Dynamic network data are becoming increasingly available, for example, social networks representing social interactions over time. Hence there is a need to develop a suitable methodology for the statistical analysis of networks which are conditional on time. Motivated by these dynamic networks, we provide a general framework to estimate a regression curve from a sample of networks which are conditional on a set of Euclidean covariates. In this framework, networks are identified by their graph Laplacian matrices, for which metrics, embeddings, tangent spaces, and a projection from Euclidean space to the space of graph Laplacians are defined. We develop an adapted Nadaraya-Watson estimator for the graph Laplacian matrices and show this has uniform weak consistency for estimation using Euclidean and power Euclidean metrics. The methodology is applied to the Enron email corpus to model smooth trends in monthly networks and highlight anomalous networks.

E0941: **Simultaneous estimation and registration of sparse, fragmented and noisy functional data**
Presenter: **Sebastian Kurtek**, The Ohio State University, United States

In many applications, smooth processes generate data that is recorded under a variety of observational regimes, including dense sampling and sparse or fragmented observations that are often contaminated with an error. The statistical goal of registering and estimating the underlying functions from discrete observations has thus far been mainly approached sequentially without formal uncertainty propagation, or in an application-specific manner by pooling information across subjects. We propose a Bayesian framework for simultaneous registration and estimation, which is flexible enough to accommodate inference on individual functions under general observational regimes. We rely on the specification of strongly informative prior models over the amplitude component of function variability using two strategies: a data-driven approach that defines an empirical basis for the amplitude subspace based on training data, and a shape-restricted approach when the relative location and number of extrema is well-understood. The proposed methods build on the elastic functional data analysis framework to separately model amplitude and phase variability. We emphasize the importance of uncertainty quantification and visualization of these two components as they provide complementary information about the estimated functions. We validate the proposed framework using simulation studies and real applications.

E1037: **Modeling partially observed data with spatio-temporal dependence via regression with PDE penalization**
Presenter: **Eleonora Arnone**, Politecnico di Milano, Italy
Co-authors: Laura Sangalli, Andrea Vicini

A spatio-temporal regression technique with differential regularization is studied. Through this technique, we analyze partially observed functional data with spatio-temporal dependence. We can think of spatio-temporal data as curves sampled in scattered spatial locations or surfaces observed at some time instants. The observability of these data can be of various types. For example, in the simplest case, the datum is observed uniformly in space and time. In other cases, the missing data are clustered in sub-regions. For example, we can have that, for a fixed spatial location, the corresponding curve is not observed in a long temporal interval. Vice versa, it can be the case that for a fixed time instant, the corresponding surface is not observed in a large area of the spatial domain. We focus on the partial observability characteristics of the data, and we study the proposed methodology on simulated data corresponding to different observability patterns. The methodology is suited for dealing with complicated spatial domains or signals that exhibit complex local features. Finally, we consider an application to the lake surface water temperature data. These data have a high proportion of missing values in a complex pattern, and the reconstruction of the complete signal is of great importance for climate studies.

EO211 Room R13 ADVANCES IN SPORTS Chair: David Clarke

E0799: **Applying Bayesian inference to the impulse-response model of athletic training and performance**
Presenter: **David Clarke**, Simon Fraser University, Canada
Co-authors: Kangyi Peng, Ryan Brodie, Tim Swartz

The impulse-response (IR) model describes the relationship between athlete training history and performance. It takes as input daily training loads and fits them to past performance data. The model features five adjustable parameters and two derived parameters. Despite some past successes, IR models are often poorly estimated. We describe a novel Bayesian inference approach to estimate the IR model. We discuss the basis of informative priors and justify the assumption that performance conforms to a multivariate normal distribution. Markov chain Monte Carlo simulation (MCMC), via Gibbs sampling, was used to sample the posterior distributions. The method was applied to data from an international-class middle-distance runner, for which training was quantified as individualized training impulse and performance as IAAF points achieved in a sanctioned race. The inference procedure produced well-constrained estimates of the five adjustable parameters, but the posterior interval widths of the derived parameters were too wide to make reliable training optimization decisions. We conclude by discussing modifications to our approach that could further improve the Bayesian inference of the IR model.

E1135: **The impact of Four Factors on a basketball team success: An approach with model-based recursive partitioning**
Presenter: **Manlio Migliorati**, University of Brescia, Italy
Co-authors: Marica Manisera, Paola Zuccolotto

According to some basketball experts, statistics are killing basketball. In our opinion, they are right, if statistics reduce the game to numbers that are not truly able to describe it. Instead, sound statistical methods start from those statistics as the input data, and appropriately elaborate and transform them into useful information to support technical experts. In the last decades, publications on statistics in basketball have multiplied and tried to answer different research questions: forecasting the outcomes of a game, analysing players performance, identifying optimal game strategies. We study the evolution of the weight of the Oliver Four Factors as determinants of the probability of winning a basketball game, using data from 19138 matches of 16 NBA regular seasons (from 2004-2005 to 2019-2020). Four Factors identify team strengths and weaknesses: shooting, turnovers, rebounding and free throws. Intending to investigate the role of each factor in determining a team success, we applied the MOB algorithm for model-based recursive partitioning that, instead of fitting one global model to the entire dataset, estimates local models on clusters of matches that are defined according to a learning algorithm based on recursive partitioning.

EO512 Room R14 NON-STANDARD STATISTICS ON COMPLEX DATA Chair: Cecile Durot

E0561: **Isotonic regression meets LASSO**
Presenter: **Matey Neykov**, Carnegie Mellon University, United States

A two-step procedure is considered for monotone increasing and smooth additive single index models with Gaussian designs. The proposed procedure is simple, easy to implement with existing software, and consists of consecutively applying LASSO and isotonic regression. Aside from formalizing this procedure, we provide theoretical guarantees regarding its performance: 1) we show that our procedure controls the in-sample squared error; 2) we demonstrate that one can use the procedure for predicting new observations, by showing that the absolute prediction error can be controlled with high-probability. Our bounds show a tradeoff of two rates: the minimax rate for estimating the high dimensional quadratic

loss, and the minimax nonparametric rate for estimating a monotone increasing function. Time permitting we may also consider applying the same procedure to binary single index models with Gaussian design.

E0687: **Some rates of convergence in unlinked monotone regression**
Presenter: **Fadoua Balabdaoui**, ETH Zurich, Switzerland

The so-called univariate unlinked regression is considered when the unknown regression curve is monotone. In standard monotone regression, one observes a pair (X,Y) where a response Y is linked to a covariate X through the model $Y = m_0(X) + \varepsilon$, with m_0 the (unknown) monotone regression function and ε the unobserved error (assumed to be independent of X). In the unlinked regression setting one gets only to observe a vector of realizations from both the response Y and from the covariate X where now Y is only known to have the same distribution as $m_0(X) + \varepsilon$. Despite this, it is actually still possible to derive a consistent non-parametric estimator of m0 under the assumption of monotonicity ofm0 and knowledge of the distribution of the noise. We establish an upper bound on the rate of convergence of such an estimator under minimal assumptions on the distribution of the covariate X. We discuss extensions to the case in which the distribution of the noise is unknown. We develop a gradient-descent-based algorithm for its computation, and we demonstrate its use on synthetic data.

E1028: **Optimal linear discriminators for the discrete choice model in growing dimensions**
Presenter: **Debarghya Mukherjee**, University of Michigan, United States
Co-authors: Moulinath Banerjee, Yaacov Ritov

Manski's celebrated maximum score estimator for the discrete choice model, which is an optimal linear discriminator, has been the focus of much investigation in both the econometrics and statistics literature. Still, its behavior under growing dimension scenarios largely remains unknown. That gap is addressed. Two different cases are considered: p grows with n, but at a slow rate, i.e. $p/n \to 0$; and $p >> n$ (fast growth). In the binary response model, we recast Manski's score estimation as an empirical risk minimization for a classification problem. We derive the l_2 rate of convergence of the score estimator under a transition condition in terms of our margin parameter that calibrates the level of difficulty of the estimation problem. We also establish upper and lower bounds for the minimax l_2 error in the binary choice model that differ by a logarithmic factor and construct a minimax-optimal estimator in the slow growth regime. Some extensions to the general case – the multinomial response model – are also considered. Last but not least, we use a variety of learning algorithms to compute the maximum score estimator in growing dimensions.

EO528 Room R15 TOPICS IN CAUSAL INFERENCE: SELECTION BIAS AND SENSITIVITY ANALYSIS Chair: Ingeborg Waernbaum

E0634: **Simple sensitivity analysis for selection bias using bounds**
Presenter: **Louisa Smith**, Harvard T.H. Chan School of Public Health, United States

When epidemiologic studies are conducted in a subset of the population, selection bias can threaten the validity of causal inference. This bias can occur whether or not that selected population is the target population, and can occur even in the absence of exposure-outcome confounding. However, it is often difficult to quantify the extent of selection bias, and sensitivity analysis can be challenging to undertake and to understand. We demonstrate that the magnitude of the bias due to selection can be bounded by expressions defined by parameters characterizing the relationships between unmeasured factor(s) responsible for the bias and the measured variables. No functional form assumptions are necessary for those unmeasured factors. Using knowledge about the selection mechanism, researchers can account for the possible extent of selection bias by specifying the size of the parameters in the bounds. We also show that the bounds, which differ depending on the target population, result in summary measures that can be used to calculate the minimum magnitude of the parameters required to shift a risk ratio or risk difference to the null. A summary measure can be used as a simple sensitivity analysis to determine the overall strength of the selection that would be necessary to explain away a result. When researchers are willing to make assumptions or have knowledge about the selection mechanism, the bounds and summary measures can be further simplified.

E0481: **The E-value: Methodological clarifications**
Presenter: **Arvid Sjolander**, Karolinska Institute, Sweden

Most epidemiological studies aim to estimate the causal effect of a particular exposure on a particular outcome. Recently, the so-called "E-value" was proposed, as a tool to assess the sensitivity of obtained estimates to unmeasured confounding. The E-value has quickly become very popular, and is today often recommended. We will clarify some important methodological points regarding the sensitivity analysis, on which the E-value is based. Specifically, we will derive the feasible range for the underlying sensitivity analysis parameters, and we will discuss when and why the sensitivity analysis can be expected to be conservative.

E0792: **A potential outcomes approach to selection bias**
Presenter: **Eben Kenah**, The Ohio State University, United States

Selection bias occurs when the association between exposure and disease in the study population differs from that in the population eligible for inclusion. Along with confounding, it is one of the fundamental threats to the validity of epidemiologic research. We propose a definition of selection bias in terms of potential outcomes. This approach generalizes a previous structural approach which defines selection bias as a distortion of the exposure-disease association that is caused by conditioning on a collider. Both approaches agree in all situations where the structural approach identifies selection bias, but the potential outcomes approach identifies selection bias in situations where the earlier approach does not. Selection bias defined by potential outcomes can involve a collider at exposure, a collider at disease, or no collider at all. This broader definition of selection bias does not depend on the parameterization of the association between exposure and disease, so it can be analyzed using nonparametric single-world intervention graphs (SWIGs) both under the null hypothesis and away from it. It provides a more nuanced interpretation of the role of randomization in clinical trials, simplifies the analysis of matched studies and case-cohort studies, and distinguishes more clearly between the estimation of causal effects within the study population and generalization to the eligible population.

EO502 Room R16 RECENT ADVANCES IN CAUSAL INFERENCE Chair: Yeying Zhu

E0501: **Covariate balancing for robust estimation of causal effects of general treatment regimes**
Presenter: **Xavier de Luna**, Umea University, Sweden

Novel robust estimators for categorical and continuous treatment regimes are proposed by using a strategy based on covariate balancing propensity score and inverse probability weighting. The resulting estimators are shown to be consistent and asymptotically normal for causal contrasts of interest, either when the covariate dependent treatment assignment model is correctly specified, or when the correct set of bases for the outcome models in the space spanned by the covariates has been chosen, and the assignment model is sufficiently rich. For the categorical treatment regime case, we show that the estimator attains the semiparametric efficiency bound when all both models are correctly specified. For the continuous case, the causal contrasts of interest are functions. The latter are not parametrized and the estimators proposed are shown to have bias and variance of the classical nonparametric rate. Asymptotic results are complemented with simulations illustrating the finite sample properties. Our analysis of a data set suggests a nonlinear causal effect of BMI on the decline in self-reported health.

E0685: **Treat thy neighbour: Precision medicine in networks**
Presenter: **Michael Wallace**, University of Waterloo, Canada

Precision medicine describes the practice of tailoring treatment decisions to patient-level characteristics such as symptom severity, age, or prior medication. This may be formalized through dynamic treatment regimes: sequences of treatment decision rules that take patient information as

input and output treatment recommendations. The dynamic treatment regime and precision medicine literatures typically make the assumption of no interference: that one patient's treatment does not affect the outcome of another patient. This assumption is often violated, such as in the study of infectious diseases where treating one patient may not only lower their risk of infection, but by extension the risk of infection for those they come into contact with. We discuss the implications of interference in the context of dynamic treatment regimes, demonstrate how it may be accounted for in analysis, and highlight some of the challenges associated with the order in which treatment decisions are made within a network structure.

E0827: **Nonparametric inverse probability weighted estimators based on the highly adaptive lasso**

Presenter: **Ashkan Ertefaie**, University of Rochester, United States

Co-authors: Mark van der Laan, Nima Hejazi

Inverse probability weighted estimators are the oldest and potentially most commonly used class of procedures for the estimation of causal effects. By adjusting for selection biases via a weighting mechanism, these procedures estimate an effect of interest by constructing a pseudo-population in which selection biases are eliminated. Despite their ease of use, these estimators require the correct specification of a model for the weighting mechanism, are known to be inefficient and suffer from the curse of dimensionality. We propose a class of nonparametric inverse probability-weighted estimators in which the weighting mechanism is estimated via undersmoothing of the highly adaptive lasso. We demonstrate that our estimators are asymptotically linear with variance converging to the nonparametric efficiency bound. Unlike doubly robust estimators, our procedures require neither derivation of the efficient influence function nor specification of the conditional outcome model. Our theoretical developments have broad implications for the construction of efficient inverse probability-weighted estimators in large statistical models and a variety of problem settings. We assess the practical performance of our estimators in simulation studies and demonstrate the use of our proposed methodology with data from a large-scale epidemiologic study.

EO526 Room R17 HIGH-DIMENSIONAL INFERENCE FOR COMPLEX PROBLEMS Chair: Yumou Qiu

E1124: **Clustering with lat semiparametric mixture models**

Presenter: **Wen Zhou**, Colorado State University, United States

Co-authors: Lyuou Zhang, Hui Zou, Lulu Wang

Model-based clustering is one of the fundamental statistical approaches in unsupervised learning and has a wide range of applications. While modeling the clusters by a mixture distribution is concise and easy to implement, the traditional distributional assumptions such as the Gaussianity or other parametric forms are stringent in practice and not always realistic to verify. Existing efforts on relaxing such assumptions, on the other hand, are mostly algorithmic without any guarantees on the performance. We introduce a novel latent semiparametric mixture model to facilitate clustering data without imposing any direct distributional assumptions on data. Specifically, the model only assumes that the observations are generated from some unknown monotone transformations of latent variables governed by a Gaussian mixture. The nontrivial identifiability of the proposed model due to the unknown transformations is carefully studied. For implementation, we introduce an alternating maximization procedure based on the EM algorithm and scrupulously investigate its convergence using finite-sample analysis. An interesting transition phenomenon on the convergence of the proposed algorithm, which is due to the presence of the unknown transformations, is explored and guides the execution of the algorithm. This observation also leads to the rate of convergence for the excess mis-clustering error of our method compared to the traditional results.

E1126: **Sharp optimality for high dimensional covariance testing**

Presenter: **Yumou Qiu**, Iowa State University, United States

The theoretical limit of testing a high-dimensional covariance being diagonal is developed by deriving the sharp detection boundary as a function of signal proportion and signal strength under alternative hypotheses. The detection boundary gives the exact minimal signal strength that can be detected by some test under the sparse and faint signal regime, which is the most challenging setting for signal detection. We develop an optimal test by multi-level thresholding that can achieve the detection boundary. The optimality means the proposed test is powerful as long as the signal strength is above the detection boundary. We establish the asymptotic distribution of the thresholding statistic under non-Gaussian data. A novel U-statistic composition is developed in conjunction with the matrix blocking and the coupling techniques to handle the complex dependence among sample covariances. We show that the existing tests are non-optimal, and the proposed tests are more powerful than those existing tests. Simulation studies are conducted to demonstrate the utility of the proposed test.

E1128: **Doubly robust semiparametric difference-in-differences estimators with high-dimensional data**

Presenter: **Jing Tao**, University of Washington, United States

A doubly robust two-stage semiparametric difference-in-difference estimator is proposed for estimating heterogeneous treatment effects with high-dimensional data. The new estimator is robust to model miss-specifications and allows for, but does not require, many more regressors than observations. The first stage allows a general set of machine learning methods to be used to estimate the propensity score. In the second stage, we derive the rates of convergence for both the parametric parameter and the unknown function under a partially linear specification for the outcome equation. We also provide bias correction procedures to allow for valid inference for the heterogeneous treatment effects. We evaluate the finite sample performance with extensive simulation studies. Additionally, a real data analysis on the effect of the Fair Minimum Wage Act on the unemployment rate is performed as an illustration of our method.

EO494 Room R18 STATISTICAL AND MACHINE LEARNING METHODOLOGY FOR MEDICAL IMAGING Chair: John Kornak

E1076: **Adaptive regularization for multi-modal brain imaging**

Presenter: **Jaroslaw Harezlak**, Indiana University School of Public Health-Bloomington, United States

Co-authors: Damian Brzyski, Kewin Paczek, Timothy Randolph, Joaquin Goni

The problem of adaptive incorporation of multi-modal brain imagining data in the multiple linear regression setting is addressed. We assume the model of the form $E[Y|X,Z] = X * beta + Z * b$, where the response variable Y corresponds to a neuropsychological outcome, X are the possible confounders, and Z are the explanatory variables (e.g. cortical thickness or area) for which the functional and structural connectivity information exists. The connectivity information is used to build the adaptive penalty terms in the regularized regression problem. The general idea of incorporating connectivity information in regularization approach via linear mixed model representation has been recently established in our prior work: ridgified Partially Empirical Eigenvectors for Regression (riPEER). We incorporate multiple sources of information, e.g. functional and structural connectivity network structure, and estimate the regression parameters with multiple penalty terms via a riPEER extension called AIMER (Adaptive Information Merging Estimator for Regression). We present an extensive simulation study testing various realistic scenarios and apply msPEER to data arising from the Human Connectome Project (HCP) study.

E1099: **The linear additive tree**

Presenter: **Efstathios Gennatas**, University of California, San Francisco, United States

Co-authors: Jerome Friedman, Eric Eaton, Charles Simone II, Lyle Ungar, Lei Xing, Gilmer Valdes

The Linear Additive Tree (LINAD) is a novel algorithm that builds highly accurate and interpretable decision trees with linear models in the terminal nodes. An extension of the Additive Tree, LINAD capitalizes on the complementary strengths of decision trees and linear models with the additive training of gradient boosting and can be considered a generalization of these three algorithms. A single LINAD is fully interpretable and rivals the performance of ensemble techniques, while an ensemble of LINADs can match or surpass current ensembles based on traditional

decision trees. The algorithm's performance is demonstrated using a collection of 72 real and synthetic publicly available datasets. Across the 72 benchmarks, LINAD ranked ahead of random forest and just behind gradient boosting. At the same time, LINAD ensembles without any tuning matched the performance of gradient boosting while using many fewer trees. Following these successful benchmarks, LINAD will be applied on magnetic resonance scans from a large cohort of children (>10k). High-dimensional neuroimaging data will be used to derive patterns of white and grey matter structure that predict neurocognitive performance. LINAD offers both accuracy and interpretability to foster discovery in basic research and provide trustworthiness in critical applications like clinical predictive modeling.

E0649: **Time-varying dynamic network model for extracting the dynamic resting-state functional connectivity**
Presenter: **Fei Jiang**, The University of California, San Francisco, United States

Dynamic resting-state functional connectivity (RSFC) is believed to reflect the intrinsic organization and network structure of brain regions. The existing methods to extract dynamic RSFCs do not adapt to different datasets. Furthermore, they are not suitable for multi-modality studies. Moreover, it is difficult to justify that the resulting dynamic RSFCs are the intrinsic features that generate the brain signals. To overcome these deficiencies, we develop a time-varying dynamic network (TVDN) framework to extract the resting-state functional connectivity from neuroimaging data. TVDN has a fully automatic parameter turning mechanism, and hence it is adaptive to different datasets. Furthermore, TVDN is easily generalizable to handle the multi-modality data. Moreover, TVDN describes the relation between RSFC and brain signals. Hence, it is easy to evaluate the method by examining whether the resulting features can reconstruct the observations. We develop the TVDN model and the estimation procedures. Furthermore, we conduct comprehensive simulations to evaluate TVDN under hypothetical settings. Finally, we apply the TVDN on both fMRI and MEG data and compare the results with the existing approaches. The results show that the TVDN is more robust to detect brain state switching in the resting state. In addition, the resulting dynamic RSFCs directly link to the signal frequency and growth/decay constant and can uncover the noiseless brain signals.

EO155 Room R19 STATISTICAL METHODS FOR IMAGING DATA ANALYSIS Chair: Mark Fiecas

E0606: **Modeling populations of networks from multi-subject neuroimaging data**
Presenter: **Subhadeep Paul**, The Ohio State University, United States

To analyze data from multi-subject experiments in neuroimaging studies, we develop a modeling framework for joint community detection in a population of networks. The proposed random effects stochastic block model facilitates the study of group differences and subject-specific variations in the community structure. The model proposes a putative mean community structure that is representative of the population under consideration but is not the community structure of any individual component network. Instead, the community memberships of nodes vary in each component network with a transition matrix, thus modeling the variation in community structure across a group of subjects. To estimate the quantities of interest, we propose two methods: a variational EM algorithm, and a non-negative matrix factorization based method called Co-OSNTF. We also develop a resampling-based hypothesis test for differences between community structure in two populations both at the whole network level and node level. The methodology is applied to data on multi-subject experiments involving schizophrenia and Tinnitus patients in two separate works.

E0237: **A Bayesian spatial model for imaging genetics**
Presenter: **Farouk Nathoo**, University of Victoria, Canada

A Bayesian bivariate spatial model is developed for multivariate regression analysis applicable to studies examining the influence of genetic variation on brain structure. A bivariate spatial process model is developed to accommodate the correlation structures typically seen in structural brain imaging MRI data. First, we allow for spatial correlation in the imaging phenotypes obtained from neighbouring regions on the same hemisphere of the brain. Second, we allow for correlation in the same phenotypes obtained from different hemispheres (left/right) of the brain. To do this we employ a proper bivariate conditional autoregressive spatial model for the errors in a bivariate spatial regression model. Two approaches are developed for Bayesian computation: (i) a mean-field variational Bayes algorithm and (ii) a Gibbs sampling algorithm. In addition to developing the spatial model and computational procedures to approximate the posterior distribution, we also incorporate Bayesian false discovery rate (FDR) procedures to select SNPs. The methodology is illustrated through a simulation study and an application to data obtained from the Alzheimer's Disease Neuroimaging Initiative study.

E0224: **Stochastic modelling for desorption in mass spectrometry imaging**
Presenter: **Xavier Loizeau**, National Physical Laboratory, United Kingdom
Co-authors: Rory Steven, Martin Metodiev, Josephine Bunch

Mass spectrometry imaging (MSI) techniques enable spatially resolved detection of thousands of molecular species from a complex sample, such as biological tissue, in a single experiment. Unfortunately, an MSI dataset does not typically provide a quantitative representation of detected species as physicochemical processes underlying MSI (desorption, and ionisation) are not achieved with equal efficiency for all species, and are highly dependent on the molecules present in any given region, limiting any link between observed signal and local chemical concentration, or quantities of matter. To address this issue, the biological sample of interest is modelled here as a point process and a quantitative imaging experiment is defined as any experiment that is a consistent estimator of the intensity field of this point process. This formulation gives additional meaning to many of the data mining techniques used in the MSI community, and insights on the link between MSI and challenges in modern statistic. In this model, relating the desorbed material in an MSI experiment to the intensity field of interest is done through an ill-posed inverse problem, with an unknown operator that must be estimated through calibration.

EO151 Room R20 STATISTICAL ANALYSIS OF MULTIPLE NETWORKS Chair: Eric Kolaczyk

E0300: **Estimation and bootstrapping for collections of low-rank networks**
Presenter: **Keith Levin**, University of Wisconsin, United States

In increasingly many settings, data sets consist of multiple samples from a population of networks, with vertices aligned across networks. For example, in neuroimaging, fMRI studies yield graphs whose vertices correspond to brain regions, which are the same across subjects. We consider the setting where a collection of networks share a low-rank mean structure but may differ in the noise structure on their edges. We introduce a weighted network average for estimating the low-rank structure under this setting, which we conjecture to be minimax. The utility of this estimate for inference is illustrated on synthetic networks and on data from an fMRI study of schizophrenia. We then turn to the problem of bootstrapping under this and related models for generating collections of low-rank networks. This problem raises interesting questions concerning how to conduct resampling in latent space network models.

E0373: **Spectral-based tests for hypothesis testing on populations of networks**
Presenter: **Lizhen Lin**, The University of Notre Dame, United States
Co-authors: Li Chen, Nathan Josephs, Jie Zhou, Eric Kolaczyk

The increasing prevalence of multiple network data in modern science and engineering calls for developments of models and theories that can deal with inference problems for populations of networks. We focus on addressing the problem of hypothesis testing for populations of networks and in particular, for differentiating distributions of two samples of networks. We consider a very general framework which allows us to perform tests on populations of large and sparse networks. We propose two spectral-based tests. The first test is based on the singular value of a generalized Wigner matrix. The asymptotic null distribution of the statistics is shown to follow the Tracy-Widom distribution as the number of nodes tends to infinity.

The test also yields asymptotic power guarantee with the power tending to one under the alternative. The second spectral-based test statistic is based on the trace of the third-order for a centered and scaled adjacency matrix, which is proved to converge to the standard normal distribution N (0, 1) as nodes number tends to infinity. The proper interplay between the number of networks and the number of nodes for each network are explored. We also prove the asymptotic power guarantee for this test. Extensive simulation studies and real data analyses demonstrate the superior performance of our procedure with competitors.

E0638: **Inference in the Fréchet regression model for object responses**
Presenter: **Alexander Petersen**, Brigham Young University, United States
Co-authors: Paromita Dubey, Hans-Georg Mueller

Linear regression is one of the most widely used fundamental tools in statistics for modeling response predictor relationships. In modern applications, one often encounters paired data where the responses are non-Euclidean objects (e.g., networks or covariance matrices) and predictors take values in $\mathbb{R}^d$. Fréchet regression is a state of the art method which is geared towards modeling regression between a metric space valued response and multivariate predictors. We propose the means to test the goodness of fit in Fréchet regression models, beginning with a test for the null hypothesis that the global Fréchet regression function does not depend on the predictors. We also extend the approach to investigate the partial effect of adding new predictors in an existing Fréchet regression model. The criteria we propose for the above tests are asymptotically degenerate under standard regularity conditions. To overcome this limitation, we propose the use of random multipliers, independent of the data, that give a non-degenerate distribution of the proposed test statistic under the null hypothesis of no regression effect. We illustrate the performance of the proposed test through multiple experiments in simulations, including samples of symmetric positive definite matrices equipped with various non-Euclidean metrics.

EO177 Room R21 SURVIVAL ANALYSIS Chair: Marialuisa Restaino

E0344: **Non-existence, non-uniqueness and potential uselessness of the NPMLE for doubly truncated data**
Presenter: **Carla Moreira**, University of Minho, Portugal
Co-authors: Jacobo de Una-Alvarez

Doubly truncated data are found in astronomy, epidemiology and survival analysis literature. They arise when each observation is confined to an interval; that is, the variable of interest is observed only when it falls within two random limits. The existing literature contains many nonparametric methods for dealing with truncated data. The nonparametric maximum likelihood estimator (NPMLE) for doubly truncated data has been developed. This estimator was obtained earlier for singly truncated data. To compute the NPMLE of the cumulative distribution function, iterative algorithms have been proposed. When analysing a particular doubly truncated dataset, non-existence or non-uniqueness of the NPMLE may occur; the NPMLE may be useless even when the iterative algorithms reach the convergence due to its huge variance. We present and analyse the age at diagnosis of Acute Coronary Syndrome dataset, in which these features concerned to the NPMLE appear. We apply the sufficient and necessary conditions introduced previously to investigate the existence and uniqueness of the NPMLE. We present a simulation study to investigate the impact of the width of the observational window in the potential uselessness of the NPMLE.

E0981: **A progressive three-state model for the restricted residual mean time**
Presenter: **Giuliana Cortese**, University of Padua, Italy
Co-authors: Davide Serafin

The residual restricted mean lifetime (RRML) is an alternative robust summary measure that captures a global profile of the survival curve over $[t, t_0]$ by the integration of the survival function. This measure describes the remaining life expectancy up to t_0 of an individual surviving up to a specific time t, and it permits to make easier and better assessments from a clinical point of view than those based on the hazard ratio from a proportional hazards model. Direct regression models for RRML have gained attention recently, and different inferential approaches have been proposed to handle both right censoring and left truncation. However, the multi-state setting has not yet been fully explored in regression models for the RRML. Therefore, a progressive three-state model is presented to handle situations where subjects may experience two different events, involving three possible states visited by a subject. For each transition, we assess the effect of covariates on the residual mean lifetime scale rather than on transition intensities. The inferential procedure underlying the proposed model, based on IPCW weights, is investigated by simulation studies. An application on cutaneous melanoma is illustrated to evaluate the effects of possible biomarkers on the residual mean lifetime to tumor recurrence and death after recurrence.

E1027: **Exploring and analysing the determinants of football coach dismissal in Italian League Serie A**
Presenter: **Mariangela Zenga**, Universita degli Studi di Milano-Bicocca -DISMEQ, Italy
Co-authors: Francesco Porro, Marialuisa Restaino, Juan Eloy Ruiz-Castro

The aim is to study the impact of a set of coach and team characteristics on performance and dismissals of managers (head coaches) in the top division of the Italian Football League during seasons 20162007 to 20192020 (Serie A). In particular, we examine the probability of coaches' survival by employing survival methods and stochastic processes to explore the effects of covariates on coach tenure length. Thus, we capture the variation across the seasons and assess the association between team/coach characteristics and coach dismissals. The set of coach characteristics includes both performance-related and non-performance variables and manager duration. Performance variables include characteristics of teams able to analyze and measure coach or/and team typical performance. Non-performance variables is a group of demographic characteristics related to managers and clubs. Data are extracted from some football websites.

EO141 Room R22 STATISTICAL MODELING OF COVID-19 PANDEMIC Chair: Rob Deardon

E0715: **Measuring community vulnerability in spatiotemporal disease mapping of COVID-19**
Presenter: **Rachel Carroll**, University of North Carolina Wilmington, United States

Community vulnerability is an important measure to consider in modeling the spread and impact of COVID-19. This spatially dependent factor can be described by several variables or reduced to a composite measure. Several well established composite measures of community vulnerability exist, including the social vulnerability index and the area deprivation index. A new, additional measure of vulnerability specifically related to COVID-19 was developed in recent months - the COVID-19 community vulnerability index. We compare these four methods of accounting for community vulnerability in the modeling of COVID-19. The statistical model used was a spatiotemporal Bayesian Poisson regression model, and models were compared to fit via the deviance information criterion and the Watanabe-Akaike information criterion. Results suggest a better model fit when including any of the vulnerability measures, and all indices generally suggest a higher risk of COVID-19 in areas with more vulnerability. The COVID-19 community vulnerability index has some benefit in terms of model fit compared to the others considered, particularly when examining COVID-19 deaths. Still, insights on specific variable relationships with COVID-19 are not accessible when using composite measures. In conclusion, it is important to adjust for community vulnerability in modeling of COVID-19, but the best measure depends on the goal of the assessment.

E0777: **Spatiotemporal transmission dynamics of COVID-19 in Nigeria**
Presenter: **Ashok Krishnamurthy**, Mount Royal University, Canada
Co-authors: Bedrich Sousedik, Maya Mueller, Agatha Ojimelukwe, Brittany Millis, Jocelyn Boegelsack, Loren Cobb

The global coronavirus pandemic (COVID-19) reached Lagos, Nigeria on February 27, 2020. Since then, COVID-19 infections have been re-

ported in the majority of Nigerian states. We present a spatial Susceptible-Exposed-Infectious-Recovered-Dead (SEIRD) compartmental model of epidemiology to capture the transmission dynamics of the spread of COVID-19 and provide insight that would support public health officials towards informed, data-driven decision making. Data assimilation is a general category of statistical tracking techniques that incorporate and adapt to real-time data as they arrive by sequential statistical estimation. Data assimilation applied to the SEIRD model receives daily aggregated epidemiological data from the Nigeria Centre for Disease Control (NCDC) and uses this data to perform corrections to the current state vector of the epidemic. In other words, it enhances the operation of the SEIRD model by periodically executing a Bayesian correction to the state vector, in a way that is, at least arguably, statistically optimal. We observe that the prediction improves as data is assimilated over time. It is essential to understand what future epidemic trends will be, as well as the effectiveness and potential impact of government disease intervention measures. Predictions for disease prevalence with and without mitigation efforts are presented via spatiotemporal disease maps.

E0761: **Emulation based likelihood approximation for spatial infectious disease models**

Presenter: **Gyanendra Pokharel**, University of Winnipeg, Canada

Mechanistic models for spatio-temporal infectious disease offer a great advantage in capturing heterogeneity in populations during an epidemic. These models quantify probabilistic outcomes regarding the risk of infection of susceptible individuals due to various susceptibility and transmissibility covariates, including their spatial distance from infectious individuals. However, such models are generally fitted in a Bayesian Markov chain Monte Carlo (MCMC) framework, which requires multiple calculations of what is often a computationally expensive likelihood function; thus, computationally prohibitive MCMC-based analysis. We propose an alternative approach, the so-called emulation technique. The model is again fitted in a Bayesian MCMC framework but replaces the computationally expensive true likelihood by the Gaussian process approximation of the likelihood function built over the design matrix constructed on a pre-defined parameter grid. We show that such method can be used to infer the model parameters and underlying characteristics of spatial disease systems and that this can be done in much more computationally efficient manner compared to the full Bayesian MCMC approach.

EO241 Room R23 RECENT ADVANCES IN BAYESIAN METHODS FOR CORRELATED DATA Chair: Garritt Page

E0600: **Posterior summaries of topic models: An example from grocery retail baskets**

Presenter: **Ioanna Manolopoulou**, University College London, United Kingdom

Co-authors: Mariflor Vega Carrasco, Mirco Musolesi

Understanding the shopping motivations behind market baskets is an important goal in the grocery retail industry. Analyzing shopping transactions demands techniques that can cope with the volume and complicated dependencies of grocery transactional data, while keeping interpretable outcomes. Latent Dirichlet Allocation (LDA) provides a suitable framework to process grocery transactions and to discover a broad representation of customers' shopping motivations. However, summarising the posterior distribution of an LDA model is challenging, because LDA is inherently a mixture model and can exhibit substantial label-switching. Moreover, even when a posterior mean is computed, a summary of corresponding uncertainty is not straightforwardly available. We introduce a clustering methodology that post-processes posterior LDA draws to summarise the entire posterior distribution and identify semantic modes represented as recurrent topics. We illustrate our methods on an example from a large UK supermarket chain.

E0632: **Bayesian nonparametric density autoregression with lag selection**

Presenter: **Matthew Heiner**, Brigham Young University, United States

Co-authors: Athanasios Kottas

The aim is to develop a Bayesian nonparametric autoregressive model applied to estimate transition densities exhibiting nonlinear lag dependence flexibly. The approach is related to Bayesian density regression using Dirichlet process mixtures, with the Markovian likelihood defined through the conditional distribution obtained from the mixture. This results in a Bayesian nonparametric extension of a mixtures-of-experts model formulation. We illustrate and explore inferences available through the base model by fitting to synthetic and real-time series. We then explore model extensions to include global and local selection among a pre-specified set of lags, and modifications to the kernel weight function to accommodate heterogeneous dynamics. We also compare transition density estimation performance for alternate configurations of the proposed model.

E0697: **Mixed models for spatially correlated data using PC priors**

Presenter: **Massimo Ventrucci**, University of Bologna, Italy

Co-authors: Maria Franco Villoria

Generalized linear mixed models (GLMM) represent a flexible tool to model environmental data which are characterized by various sources of heterogeneity, e.g. spatial or temporal correlation. The usual interpretation is that fixed effects explain the response by measuring the effect of observed covariates, while random effects account for heterogeneity due to unobserved factors. Most popular models for random effects are Gaussian conditional on some flexibility parameter (e.g. variance, correlation range), the prior specification and estimation of which represents a crucial issue in many applications. Often, random effects have a more predominant role in the analysis and are used for explanatory purposes rather than as tools to capture residual structure; for instance, in community ecology, spatial random effects are associated to the presence of biotic interactions among species. We focus on cases where random effects reflect precise assumptions on the behavior of the phenomenon under study and propose mixed models with an intuitive prior specification. Based on the Penalized Complexity (PC) prior framework, we discuss solutions to build priors for variance parameters while achieving intuitive control on the flexibility of the random effects. We illustrate the use of the proposed priors using environmental case studies, with particular emphasis on spatially correlated data.

EO580 Room R24 PLANNED AND UNPLANNED PRESENCE IN OBSERVATIONAL RESEARCH Chair: Jonathan Schildcrout

E0592: **Informative presence bias in electronic health records**

Presenter: **Glen McGee**, University of Waterloo, Canada

Co-authors: Sebastien Haneuse, Brent Coull, Ran Rotem

Electronic health records (EHRs) offer unprecedented opportunities to answer epidemiological questions. However, unlike in ordinary cohort studies or randomized trials, EHR data are collected somewhat idiosyncratically. In particular, patients who have more contact with the medical system have more opportunities to receive diagnoses, which are then recorded in their EHRs. The goal is to clarify the nature and scope of this phenomenon, known as informative presence, which can cause bias if not accounted for. Whereas previous work has introduced this as an instance of confounding, we instead frame it in the context of misclassification. As a consequence, we show that informative presence bias can occur more broadly than previously thought and that covariate adjustment may not be fully correct for bias. Motivated by a study of autism spectrum disorder, we report on a comprehensive series of simulations to shed light on when to expect informative presence bias and how to mitigate it.

E0654: **Two-wave outcome-dependent sampling designs with applications to longitudinal binary data**

Presenter: **Ran Tao**, Vanderbilt University Medical Center, United States

Outcome-dependent sampling (ODS) designs are useful when resource constraints prohibit expensive exposure ascertainment on all study subjects. One class of ODS designs for longitudinal binary data stratifies subjects into three strata according to those who experience the event at none, some, or all follow-up times. For time-varying covariate effects, exclusively selecting subjects with response variation can yield highly efficient estimates. However, if interest lies in the association of a time-invariant covariate or the joint associations of time-varying and time-invariant covariates with the outcome, then the optimal design is unknown. Therefore, we propose a class of two-wave two-phase ODS designs for longitudinal binary data.

We split the second-phase sample selection into two waves, between which an interim design evaluation analysis is conducted. The interim design evaluation analysis uses first-wave data to conduct a simulation-based search for the optimal second-wave design that will improve the likelihood of study success. We believe the proposed designs can be useful in settings where 1) the second-phase sample size is fixed, and one must tailor relative sampling proportions among the strata to maximize estimation efficiency, or 2) the relative sampling proportions are fixed, and one must tailor the sample size to achieve the desired precision.

E0639: **Cluster-based outcome-dependent sampling in resource-limited settings: Inference in small-samples**
Presenter: **Sebastien Haneuse**, Harvard TH Chan School of Public Health, United States
Co-authors: Sara Sauer, Bethany Hedt-Gauthier, Claudia Rivera-Rodriguez

Outcome-dependent sampling is an indispensable tool for carrying out cost-efficient research in resource-limited settings. One such sampling scheme is a cluster-based design where clusters of individuals (e.g. clinics) are selected, in part at least, on the basis of the outcome rate of the individuals. For a given dataset collected via a cluster-based outcome-dependent sampling scheme, it has previously been proposed to perform estimation for a marginal model using inverse-probability-weighted generalized estimating equations, where the cluster-specific weights are the inverse probability of the clinics' inclusion in the sample. We provide a detailed treatment of the asymptotic properties of this estimator, together with an explicit expression for the asymptotic variance and a corresponding estimator. Furthermore, motivated by a study we conducted in Rwanda, we provide expressions for small- sample bias corrections to both the point estimates and the standard error estimates. Through simulation, we show that applying these corrections when the number of clusters is small generally reduces the bias in the point estimates, and results in closer to nominal coverage. The proposed methods are illustrated using data from 18 health centers in Rwanda, collected via a cluster-based outcome-dependent sampling scheme, with the goal of examining risk factors for low birth weight.

EO257 Room R25 RECENT ADVANCES IN ESTIMATION THEORY Chair: Gourab Mukherjee

E0302: **Estimation in tensor Ising models**
Presenter: **Bhaswar Bhattacharya**, University of Pennsylvania, United States
Co-authors: Somabha Mukherjee, Jaesung Son

The p-tensor Ising model is a one-parameter discrete exponential family for modeling dependent binary data, where the sufficient statistic is a multi-linear form of degree $p \geq 2$. This is a natural generalization of the matrix Ising model that provides a convenient mathematical framework for capturing higher-order dependencies in complex relational data. We will discuss the problem of estimating the natural parameter of the p-tensor Ising model given a single sample from the distribution on N nodes. Our estimate is based on the maximum pseudo-likelihood (MPL) method, which provides a computationally efficient algorithm for estimating the parameter that avoids computing the intractable partition function. General conditions under which the MPL estimate is $\sqrt{N}$-consistent will be presented. Our conditions are robust enough to handle a variety of commonly used tensor Ising models, including models where the rate of estimation undergoes a phase transition, such as the well-known stochastic block model. Finally, we will discuss the precise fluctuations of the MPL estimate in the special case of the Curie-Weiss model. The MPL estimate saturates the Cramer-Rao lower bound at all points above the estimation threshold, that is, the MPL estimate incurs no loss in asymptotic efficiency, even though it is obtained by minimizing only an approximation of the true likelihood function for computational tractability.

E0528: **Optimal shrinkage estimation of predictive densities under alpha–divergences**
Presenter: **Gourab Mukherjee**, University of Southern California, United States

The problem of estimating the predictive density in a heteroskedastic Gaussian model under general divergence loss is considered. Based on a conjugate hierarchical set-up, we consider generic classes of shrinkage predictive densities that are governed by location and scale hyper-parameters. For any alpha-divergence loss, we propose a risk-estimation based methodology for tuning these shrinkage hyper-parameters. The proposed predictive density estimators enjoy optimal asymptotic risk properties that are in concordance with the optimal shrinkage calibration point estimation results. These alpha-divergence risk optimality properties of the proposed predictors are not shared by empirical Bayes predictive density estimators that are calibrated by traditional methods such as maximum likelihood and method of moments. We conduct several numerical studies to compare the non-asymptotic performance of our proposed predictive density estimators with other competing methods and obtain encouraging results. We demonstrate the applicability of these shrunken predictive densities for analyzing protein expression data from virus-infected cell samples.

E0552: **A general framework for empirical Bayes estimation in the discrete linear exponential family**
Presenter: **Trambak Banerjee**, University of Kansas, United States
Co-authors: Gourab Mukherjee, Wenguang Sun

A Nonparametric Empirical Bayes (NEB) framework is developed for compound estimation in the discrete linear exponential family, which includes a wide class of discrete distributions frequently arising from modern big data applications. We propose to directly estimate the Bayes shrinkage factor in the generalized Robbins' formula via solving a convex program, which is carefully developed based on an RKHS representation of the Stein's discrepancy measure. The new NEB estimation framework is flexible for incorporating various structural constraints into the data-driven rule, and provides a unified approach to compound estimation with both regular and scaled squared error losses. We develop theory to show that the class of NEB estimators enjoys strong asymptotic properties. Comprehensive simulation studies, as well as analyses of real data examples, are carried out to demonstrate the superiority of the NEB estimator over competing methods.

CO273 Room R02 TIME SERIES AND FORECASTING Chair: Robert Kunst

C0921: **Forecasting aggregate household consumption and aggregate income: A simulation-based model selection approach**
Presenter: **Robert Kunst**, Institute for Advanced Studies, Austria
Co-authors: Adusei Jumah

Household consumption and disposable income provide a role-model example for error correction. On given national-accounts data, we explore whether and to what degree the cointegration properties benefit forecasting. It evolves that statistical evidence on cointegration is not always equivalent to better forecasting properties by the implied cointegrating structure. The exercise is conducted in the framework of simulation-based forecast-model selection. Apart from prediction by competing specifications to be selected from a small choice set, we also explore forecast combinations based on Bates-Granger weights constructed from a continuum in the same framework. The simulation-based method explicitly permits letting the forecast model choice depend on the intended time horizon of the forecast. The simulation-based approach permits the determination of the sample size, beyond which the more sophisticated model dominates with regard to its forecasting properties. The forecast combination experiments indicate a window of opportunity at a specific horizon, whereas pure strategies dominate at smaller and larger horizons.

C1014: **Daily, weekly, monthly, and quarterly hotel room demand forecasts for Vienna across hotel classes**
Presenter: **Ulrich Gunter**, MODUL University Vienna, Austria

Daily data for Vienna is employed which have been made available by the STR Share Center over the period January 1, 2010, to January 31, 2020, for the hotel classes all, luxury, upper upscale, upscale, upper midscale, midscale, and economy. The forecast variable of interest is hotel room demand (i.e., the number of rooms sold per day). As single forecast models, (1) seasonal naive, (2) ETS, (3) SARIMA, (4) Seasonal Neural Network Autoregressive (SNNAR), as well as (5) SNNAR with an additional Regressor (REG-SNNAR) are employed. As additional regressor in the REG-SNNAR model, seasonal naive forecasts of the inflation-adjusted Average Daily Rate (ADR) in euros are used. Forecast evaluation is

carried out for forecast horizons $h = 1, 7, 30$, and 90 days ahead based on rolling estimation windows (i.e., a form of time series cross-validation). As forecast combination techniques, (a) simple mean, (b) simple median, (c) least-squares weights, (d) MSE weights, and (e) MSE ranks are calculated. Based on preliminary results and a few exceptions notwithstanding, combined forecasts based on MSE (or Bates-Granger) weights and MSE ranks generally provide the highest level of forecast accuracy.

C0630: **Sparse structures with LASSO through principal components: Forecasting GDP components in the short-run**
Presenter: **Saulius Jokubaitis**, Vilnius University, Lithuania
Co-authors: Dmitrij Celov, Remigijus Leipus

The aim is to examine the use of sparse methods to forecast the real, in the chain-linked volume sense, expenditure components of the US and EU GDP in the short-run sooner than the national institutions of statistics officially release the data. We estimate current quarter nowcasts along with 1- and 2-quarter forecasts by bridging quarterly data with available monthly information announced with a much smaller delay. We solve the high-dimensionality problem of the monthly dataset by assuming sparse structures of leading indicators, capable of adequately explaining the dynamics of analyzed data. For variable selection and estimation of the forecasts, we use the sparse methods LASSO together with its recent modifications: Adaptive LASSO, Relaxed LASSO, Square Root LASSO and the Fast Best Subset Selection. We propose an adjustment that combines LASSO cases with principal components analysis that deemed to improve the forecasting performance. We evaluate forecasting performance conducting pseudo-real-time experiments for gross fixed capital formation, private consumption, imports and exports over the sample of 2005-2019. The performance is compared with benchmark ARMA and factor models. The main results suggest that sparse methods can outperform the benchmarks and to identify reasonable subsets of explanatory variables. The proposed LASSO-PC modification show further improvement in forecast accuracy.

CO095 Room R04 SENTIMENTS, UNCERTAINTY AND MACHINE LEARNING Chair: Svetlana Makarova

C0471: **Economic policy uncertainty in China since 1949: The view from mainland newspapers**
Presenter: **Xuguang Simon Sheng**, American University, United States
Co-authors: Steve Davis, Dingqian Liu

The economic policy uncertainty (EPU) in China since 1949 is quantified as filtered through the lens of two leading mainland newspapers. We use scaled frequency counts of newspaper articles that contain selected terms to quantify EPU. We rely on natural language processing tools to help select policy-relevant terms. Our evidence suggests that mainland newspapers yield a reasonable proxy for EPU in China since the mid-1990s and possibly earlier. Our index is highly elevated during the Korean War, rises sharply in 1979 amidst tensions over market-based reforms and responds to many other domestic and foreign developments that include Ronald Reagan's election as U.S. President in 1980, political battles over the role of market forces in 1986-1987, German reunification in 1990, the Global Financial Crisis of 2008-2009, and, especially, rising trade policy tensions in 2017-2018. In VAR models fit data since 1992, surprise increases in our EPU index foreshadow deteriorations in Chinas economic performance. Our trade policy uncertainty index for China skyrockets in 2018. Contemporaneously, Chinese firms with high sales to the U.S. saw large negative equity returns and large increases in return volatilities relative to other Chinese firms.

C0825: **Economic policy uncertainty and COVID-19 pandemic media coverage: The machine learning approach**
Presenter: **Svetlana Makarova**, University College London, United Kingdom
Co-authors: Wojciech Charemza, Krzysztof Rybinski

New uncertainty measures are proposed, which directly incorporate the effects of media coverage of the pandemic. We hypothesise that the excessive media coverage of the COVID-19 pandemic in 2020, crowded out other news; in particular, these related to economic policy uncertainty. As a consequence, the economic policy uncertainty (EPU) measures constructed from the frequencies of the appearance of the uncertainty-related terms in the media become biased downward, even though the pandemic reporting actually increased uncertainty. To show this, we construct health-augmented (HEPU) uncertainty indices for five countries: Belarus, Kazakhstan, Poland, Russia and Ukraine. These countries conducted, in the first half of 2020, different anti-pandemic policies and their media exhibited different approaches towards the pandemic coverage. We apply Word2vec to define cosine-similar words to distinguish health-related, economic, policy and uncertainty terms. For topics identification, we use the unsupervised learning technique based on the Latent Dirichlet Allocation method. The results confirm the impact the reporting of the pandemic had on the development of the crowding-out effect. We also evaluate the resulting bias in the EPU indices. The findings vary across the analysed countries, which is explained to the different styles of media coverage of the pandemic.

C0734: **Financial market uncertainty in the US: Measurement, trends, and effects**
Presenter: **Svatopluk Kapounek**, Mendel University in Brno, Czech Republic

Based on a textual analysis of more than 100 million articles published from 1885-2017 in 11 major US newspapers, we develop a new monthly index of financial market uncertainty as well as uncertainty subindexes for banks and stock, bond and money markets. We find that spikes in the uncertainty indexes correspond to major financial, political and policy events. The subindexes show distinct patterns, suggesting the benefits of measuring uncertainty at a more disaggregated level. The regression results show that financial uncertainty not only affects several macroeconomic and financial variables but that the reaction to financial uncertainty shocks is stronger and faster during recessions consistent with finance uncertainty multiplier theory.

CC808 Room R08 CONTRIBUTIONS IN MACROECONOMETRICS Chair: Mariarosaria Comunale

C0782: **Solving the unobserved components puzzle: A fractional approach to measuring the business cycle**
Presenter: **Tobias Hartl**, University of Regensburg, Germany
Co-authors: Rolf Tschernig, Enzo Weber

Measures for the business cycle obtained from trend-cycle decompositions are puzzling, as they often are noisy, at odds with the NBER chronology, and not well in line with economic theory. We argue that these results are driven by the neglect of fractionally integrated trends in log US real GDP. To account for fractional integration, we develop a generalization of trend-cycle decompositions that avoids prior assumptions about the long-run dynamic characteristics and treats the integration order as a random variable. The integration order is jointly estimated with the other model parameters via a quasi maximum likelihood estimator that is shown to be consistent and asymptotically normal. In addition, single-step estimators for the latent components that are identical to the Kalman filter and smoother but computationally superior are derived. We find that log US real GDP is integrated of order around 1.3, the resulting trend-cycle decomposition is in line with the NBER chronology, and the model well explains the puzzling results in the literature that result from model misspecification.

C1083: **Joint Bayesian inference about impulse responses in VAR models**
Presenter: **Atsushi Inoue**, Vanderbilt, United States

Structural VAR models are routinely estimated by Bayesian methods. Several recent studies have voiced concerns about the common use of posterior median (or mean) response functions in applied VAR analysis. We show that these response functions can be misleading because in empirically relevant settings there need not exist a posterior draw for the impulse response function that matches the posterior median or mean response function, even as the number of posterior draws approaches infinity. As a result, the use of these summary statistics may distort the shape of the impulse response function, which is of foremost interest in applied work. The same concern applies to error bands based on the upper and lower quantiles of the marginal posterior distributions of the impulse responses. In addition, these error bands fail to capture the full uncertainty about the estimates of the structural impulse responses. In response to these concerns, we propose new estimators of impulse response functions that

are consistent with Bayesian statistical decision theory, that respect the dynamics of the impulse response functions and that are easy to implement. We also propose joint credible sets for these estimators derived under the same loss function.

C0246: **Investor sentiment and global economic conditions**

Presenter: **Miguel Herculano**, University of Freiburg, Germany

Investor sentiment is measured at both global and local levels as the common component of pricing errors investors make when valuing stocks. Investor sentiment and macroeconomic factors are jointly modelled within a hierarchical dynamic factor model allowing for time-varying parameters and stochastic volatility. We extend existing methods to enable estimation of the model with the prescribed hierarchy which permits cross-country analysis. The approach allows us to control for macroeconomic conditions that may contaminate investor sentiment indices. We find that global investor sentiment is a key driving force behind domestic sentiment and global economic conditions.

CG028 Room R06 CONTRIBUTIONS IN APPLIED ECONOMETRICS I **Chair: Giorgio Primiceri**

C0230: **Optimizing credit gaps for predicting financial crises: Modelling choices and tradeoffs**

Presenter: **Mohammad Jahan-Parvar**, Federal Reserve Baord of Governors, United States

Co-authors: Daniel Beltran, Fiona Paine

The purpose is to bridge the academic and policy debates on the role of credit gaps for predicting financial crises, by integrating the modelling choices associated with trend-cycle decomposition methods into the design of crises early warning indicators (EWIs). We evaluate how the performance of EWIs is influenced by the choice of trend-cycle decomposition methods for constructing credit gaps (including the smoothness of the underlying trend), and by the policymaker's preference over false positives and false negatives. For the most common trend-cycle decomposition methods used to recover credit gaps, we find that optimally smoothing the trend improves the tradeoff between false positives and false negatives of the resulting EWIs, and thus their out-of-sample performance. The out-of-sample performance improves further once we consider a preference for robustness of the credit gap estimates to the arrival of new information.

C0387: **Estimation and testing for common breaks in interactive effects panels**

Presenter: **Yiannis Karavias**, University of Birmingham, United Kingdom

Co-authors: Joakim Westerlund

Dealing with structural breaks is an important step in most, if not all, empirical economic research. This is particularly true in panel data comprised of many cross-sectional units, such as individuals, firms or countries, which are all affected by major economic events. The worry is that if left unattended, existing breaks will manifest themselves as omitted variables, leading to inconsistent estimates of the model parameters. It is therefore important to know if and when a structural break has occurred. Of course, such knowledge is rarely available in practice, which means that it has to be inferred from the data. We need to be able to test if there is in fact a break present and, if it is, to infer the date of the break; that is, we need a break detection toolbox. The toolbox should apply to different sized panels, easy to implement and robust to general forms of unobserved heterogeneity. This last demand is potentially significant because unattended heterogeneity can be mistaken for omitted breaks. The purpose of the present paper is to develop a toolbox that meets the above list of demands.

C1108: **Fiscal spillovers: The case of us corporate and personal income taxes**

Presenter: **Daniela Hauser**, Bank of Canada, Canada

Co-authors: Romanos Priftis, Madeline Hanson

The aim is to extend the narrative identification of unanticipated tax changes in the US to 2019Q4, and to empirically assess the propagation of corporate and personal income tax shocks to a panel of the US main trading partners. A cut in both taxes leads to short-run expansions domestically, but their spillover effects differ markedly. A cut in corporate taxes produces negative spillovers, indicating that contractionary effects associated with the reallocation of investment and jobs of multinational firms outweigh potential positive effects through increased demand for country-specific goods through trade. The spillover effects of lower personal income taxes are more heterogeneous across countries, but positive.

Sunday 20.12.2020 08:45 - 10:50 Parallel Session G – CFE-CMStatistics

EO373 Room R12 SOME RECENT ADVANCES IN MIXTURE AND CLUSTER ANALYSES Chair: Geoffrey McLachlan

E0314: **Clustering of a directed graph: Bipartite clustering or not**
Presenter: **Christine Keribin**, INRIA - Paris-Saclay University, France
The Stochastic Block Model (SBM) is commonly used not only for the clustering of undirected graphs but also for the clustering of directed graphs too. This choice should be discussed because this model makes no difference between the clusters of source and target nodes. That is why the use of the Latent Block Model (LBM), building two different clusterings for the source and target nodes, could be interesting for directed graph clustering. We will analyze and discuss (structure, inference, model selection) through simulated data the comparison between SBM and LBM for directed graph clustering, and propose a methodology which will be applied on real data sets.

E0462: **Split sample hypothesis tests and confidence sets for mixture models**
Presenter: **Hien Nguyen**, La Trobe University, Australia
Recent work has demonstrated that one can construct hypothesis tests and confidence regions for arbitrary statistical models in a finite sample correct and estimator agnostic manner, using a split data likelihood ratio methodology. We demonstrate some interesting and useful applications of these methods in the mixture model context.

E0774: **Hidden Markov models for continuous multivariate data with missing responses and dropout**
Presenter: **Silvia Pandolfi**, University of Perugia, Italy
Co-authors: Francesco Bartolucci, Fulvia Pennoni
A Hidden Markov (HM) model is proposed for longitudinal continuous data with missing responses and dropout. These models assume the existence of an unobservable process, which follows a Markov chain with a discrete number of hidden states, affecting the distribution of the observed outcomes. In particular, we consider multivariate continuous responses that, for the same time occasion, are assumed to be correlated, according to a specific variance-covariance matrix, even conditionally on the hidden states. For the analysis of such kind of data, the presence of missing observations represents a relevant problem since dropout or non-monotone missing data patterns may occur. We propose an approach for inference with missing data by exploiting the steps of the Expectation-Maximization algorithm on the basis of suitable recursions. The resulting algorithm provides exact maximum likelihood estimates of model parameters under the missing-at-random assumption (MAR). We consider three different types of missing patters: (i) missing responses to one or more outcomes in a given time occasion; (ii) missing observation at one occasion followed by a proper evaluation in the subsequent time occasion (intermittent missing patterns); (iii) dropout, namely missing observation due to the early termination from the trial. The proposed model allows us to identify latent or unobserved clusters of units with homogeneous behavior and to track their evolution in a dynamic perspective.

E0913: **Estimating the covariance matrix of the maximum likelihood estimator under linear cluster-weighted models**
Presenter: **Gabriele Soffritti**, University of Bologna, Italy
Cluster-weighted regression constitutes an approach to regression analysis with random covariates in the presence of unobserved heterogeneity which also allows performing model-based cluster analysis. In recent years the research into this approach has been intense. However, estimating the covariance matrix of the maximum likelihood estimator is still an open issue because the expectation-maximisation algorithm usually employed to estimate parameters of cluster-weighted models does not require deriving an analytical expression for the Hessian matrix; thus, an evaluation of the covariance matrix of the maximum likelihood estimator is generally not available. An approach is developed in which information-based estimators of such a covariance matrix are obtained from the incomplete data log-likelihood of the multivariate Gaussian linear cluster-weighted model. To this end, analytical expressions for the score vector and Hessian matrix are obtained. Three estimators of the asymptotic covariance matrix of the maximum likelihood estimator, based on the score vector and Hessian matrix, are introduced. The performances of these estimators are numerically evaluated using simulated datasets in comparison with a bootstrap-based estimator; their usefulness is illustrated through a study aiming at evaluating the link between tourism flows and attendance at museums and monuments in two Italian regions.

E1065: **Continuous mixture of normal distributions for cluster analyses**
Presenter: **Sharon Lee**, University of Adelaide, Australia
The continuous mixture of normal distributions generalizes the normal distribution by scaling its mean, variance, or both with a (continuous) random variable, thereby allowing more flexible distributional shapes such as heavy tailedness and skewness. This renders them useful for modelling non-normal data. We present a selective overview of the continuous mixture of normal distributions suitable for model-based clustering. In particular, we consider the families of location mixture, scale mixture, and a location-scale mixture of normal distributions. We discuss their basic properties, important special/limiting cases, and methods for parameter estimation.

EO205 Room R13 STATISTICS FOR COMPLEX RANDOM SYSTEMS: THEORY AND PRACTICE Chair: Hiroki Masuda

E0268: **Adaptive estimation of a parabolic SPDE with a small noise**
Presenter: **Masayuki Uchida**, Osaka University, Japan
Co-authors: Yusuke Kaino
Parametric estimation is considered for a parabolic linear second order stochastic partial differential equation (SPDE) with a small dispersion parameter based on high frequency observations in time and space. Recently, the asymptotic normality of the minimum contrast estimators for the coefficient parameters of the discretely observed SPDE model on a fixed region has been shown. We first obtain the minimum contrast estimators of the diffusivity parameter and the curvature parameter in a parabolic linear SPDE with a small dispersion parameter by using the thinned data in space based on the high frequency observations. Next, the approximate coordinate process is derived from the minimum contrast estimators and the high frequency observations. The adaptive estimator of the remaining one unknown parameter in the SPDE with a small dispersion parameter is constructed by using the thinned data in time obtained from the approximate coordinate process. Moreover, we give some examples and simulation results of the estimators of the coefficient parameters in the SPDE with a small dispersion parameter.

E0270: **Transition density non-Gaussian CARMA models: Estimation and option pricing**
Presenter: **Lorenzo Mercuri**, University of Milan, Italy
It is shown how to approximate the transition density of a CARMA(p,q) model driven by means of a time-changed Brownian motion based on the Gauss-Laguerre quadrature. This approach allows us to introduce an estimation method that maximizes a likelihood function constructed using the approximated transition density. We also provide analytical formulas for the futures term structure and for option prices on futures when the underlying follows an exponential CARMA(p,q) model.

E0335: **Local asymptotic mixed normality for degenerate diffusion processes**
Presenter: **Teppei Ogihara**, University of Tokyo, Japan
Co-authors: Masaaki Fukasawa
Statistical inference is studied for diffusion processes whose diffusion coefficient is degenerate. Such models arise in the Langevin equation, which represents molecular dynamics and is also related to modeling of the stochastic volatility in finance. We show local asymptotic mixed normality

(LAMN) of the statistical models. The LAMN property is important in asymptotic statistical theory and enables us to discuss the asymptotic efficiency of estimators. The LAMN property has been studied for nondegenerate diffusion processes. We extend this result to degenerate diffusion processes by using the scheme with an L^2 regularity condition. We also show the LAMN property for partial observations of degenerate diffusion processes. This model is an extension of the LAMN results for one-dimensional integrated diffusion processes to multi-dimensional one.

E0461: **Non-adaptive estimation for a degenerate diffusion process**
Presenter: **Nakahiro Yoshida**, University of Tokyo, Japan
Co-authors: Arnaud Gloter

A multi-dimensional ergodic diffusion process specified by a system of stochastic differential equations is considered. The first component has a non-degenerate diffusion coefficient, and the second component has no diffusion coefficient. Each coefficient has an unknown vector parameter, and we estimate these parameters based on long-term high-frequency observations. While we focused on the adaptive method last year, here we discuss a non-adaptive method. The convergence rates of the diffusion parameter and the drift parameter in the non-degenerate component are the same as the usual ones, but the asymptotic variance is improved. The convergence of the estimator for the parameter in the degenerate component is much faster than the others. The non-adaptive method does not require any initial estimators and achieves the same asymptotic normality as the adaptive method.

E0533: **Asymptotic expansion formulas for diffusion processes based on the perturbation method**
Presenter: **Emanuele Guidotti**, University of Neuchatel, Switzerland
Co-authors: Nakahiro Yoshida

Asymptotic expansion formulas for arbitrary diffusion processes are provided in a form that facilitates implementation. The implementation in the R package YUIMA is presented. The user can now efficiently compute expected values of diffusion processes with theoretically any degree of accuracy. Between $10^9 - 10^10$ Monte Carlo simulations were required to distinguish between the true expected value of an Asian option payoff and the approximation computed in 0s-10s by the proposed expansion. The method finds applications in option pricing and other fields where expectations, moments, density estimation, filtering, and functionals of diffusion processes are involved.

EO548 Room R14 RECENT ADVANCES IN SPATIAL AND TIME SERIES MODELS Chair: Mattias Villani

E0419: **Deep Gaussian Markov random fields**
Presenter: **Per Siden**, Linkoping University, Sweden

Gaussian Markov random fields (GMRFs) are probabilistic graphical models widely used in spatial statistics and related fields to model dependencies over spatial structures. We establish a formal connection between GMRFs and convolutional neural networks (CNNs). Common GMRFs are special cases of a generative model where a 1-layer linear CNN gives the inverse mapping from data to latent variables. This connection allows us to generalize GMRFs to multi-layer CNN architectures, effectively increasing the order of the corresponding GMRF in a way which has favorable computational scaling. We describe how well-established tools, such as autodiff and variational inference, can be used for simple and efficient inference and learning of the deep GMRF. We demonstrate the flexibility of the proposed model and show that it outperforms the state-of-the-art on a dataset of satellite temperatures, in terms of prediction and predictive uncertainty.

E1006: **Spectral subsampling MCMC for multivariate time series**
Presenter: **Mattias Villani**, Stockholm University, Sweden
Co-authors: Matias Quiroz, Robert Kohn, Robert Salomone

Bayesian inference using Markov Chain Monte Carlo (MCMC) on large datasets has developed rapidly in recent years, particularly pseudo-marginal approaches based on efficient subsampling of conditionally independent observations. Spectral Subsampling MCMC extends the algorithms to univariate stationary time series with a large number of observations, for example, high-frequency data. The extension to multivariate stationary time series is presented.

E1110: **De-biased Whittle likelihood for time series and random fields**
Presenter: **Arthur Guillaumin**, New York University, United States
Co-authors: Adam Sykulski, Sofia Olhede

Maximum likelihood parameter estimation of time series and spatial models is often intractable for non-trivial covariance structures. We will discuss the de-biased Whittle Likelihood - a method we recently developed that can estimate time-series parameters for massive datasets using Fast Fourier transforms. The procedure is related to, but distinct from, the standard and well-known Whittle Likelihood. We will make these distinctions more clear, both from a practical and theoretical point of view. We have adapted the procedure to allow for missing data through a framework based on that of modulated time series. We have found numerous application benefits, and I will showcase these through an application to the study of Venus' topography as well as simulation studies for two and three-dimensional spatial data.

E0560: **Statistical deep IDE models for spatio-temporal forecasting**
Presenter: **Andrew Zammit Mangion**, University of Wollongong, Australia
Co-authors: Christopher Wikle

Conventional spatio-temporal statistical models are well-suited for modelling and forecasting using data collected over short time horizons. However, they are generally time-consuming to fit, and often do not realistically encapsulate temporally-varying dynamics. We tackle these two issues by using a deep convolution neural network (CNN) in a hierarchical statistical framework, where the CNN is designed to extract process dynamics from the process' most recent behaviour. Once the CNN is fitted, probabilistic forecasting can be done extremely quickly online using an ensemble Kalman filter with no requirement for repeated parameter estimation. We conduct an experiment where we train the model using 13 years of daily sea-surface temperature data in the North Atlantic Ocean. Forecasts are seen to be accurate and calibrated. A key advantage of the approach is that the CNN provides a global prior model for the dynamics that is realistic, interpretable, and computationally efficient to forecast with. We show the versatility of the approach by successfully producing 10-minute nowcasts of weather radar reflectivities in Sydney using the same model that was trained on daily sea-surface temperature data in the North Atlantic Ocean.

E1029: **Sparse spatial random graphs**
Presenter: **Francesca Panero**, University of Oxford, United Kingdom
Co-authors: Francois Caron, Judith Rousseau

A model is presented to describe spatial random graphs, exploiting the graphex setting in a Bayesian nonparametric framework that allows us flexibility and interpretable parameters. We provide several asymptotic results, namely that the model can describe both sparse and dense networks, is equipped with positive global and local clustering coefficients and can have power-law or double power-law degree distributions whose exponents are easily tuned. We also offer an efficient way to simulate from the model and perform posterior inference through an MCMC algorithm, and we show the results obtained on simulated and real data. Finally, we show that our proposal generalises several other spatial models, for example, the homogeneous random geometric graph and the hyperbolic random graph, and explain the relations with other proposals such as the scale-free percolation and the sparse latent space models.

EO121 Room R16 PIONEERING NEW FRONTIERS IN DISTRIBUTION AND MODELLING Chair: Andriette Bekker

E0416: **Joint modeling of mean and dispersion of the multivariate Laplace distribution**
Presenter: **Olcay Arslan**, Ankara University, Turkey
Co-authors: Yesim Guney, Fulya Gokalp Yavuz

An alternative robust approach is proposed to joint modeling of mean and scale covariance using multivariate Laplace distribution. The multivariate Laplace distribution has the same number of parameters as the multivariate normal distribution, but similar to the multivariate t-distribution. It is a heavy-tailed alternative distribution to the multivariate normal distribution. Since it has fewer parameters, the estimation procedure will be less complicated compared to the t-distribution. After we set the model, a modified Cholesky decomposition is adopted to factorize the dependence structure in terms of unconstrained autoregressive and scale innovation parameters. The parameter estimation is carried on using the maximum likelihood estimation method. The technique for the prediction of future responses is also investigated. The Fisher scoring algorithm and the EM algorithm are provided to compute the estimates. An extensive simulation study and a real data example are provided to demonstrate the performance of the proposed method based on Laplace distribution for jointly modeling mean and covariance.

E0568: **Parameter estimation for a Cauchy family of distributions on the sphere**
Presenter: **Shogo Kato**, Institute of Statistical Mathematics, Japan
Co-authors: Peter McCullagh

A Cauchy family of distributions on the sphere is proposed as a spherical extension of the wrapped Cauchy family on the circle. Some properties of the proposed family, especially those related to parameter estimation, are discussed. Three estimators for the spherical Cauchy family are presented, namely, a method of moments estimator, the maximum likelihood estimator, and an asymptotically efficient estimator. The method of moments estimator and the asymptotically efficient estimator are expressed in closed form. A simple algorithm is presented to estimate the maximum likelihood estimate numerically. The EM algorithm is also available for maximum likelihood estimation by transforming the spherical Cauchy family into a t-family on the Euclidean space via the stereographic projection. Asymptotic properties of the proposed estimators are considered. A simulation study is carried out to compare the estimators in terms of their performance for finite sample sizes.

E0698: **A new multivariate elliptical heavy-tailed distribution with application to allometry**
Presenter: **Antonio Punzo**, University of Catania, Italy
Co-authors: Luca Bagnato

There are several real situations where the empirical distribution of multivariate real-valued data is elliptical and with heavy tails. Many statistical models already exist that accommodate these peculiarities. This branch of literature is enriched by introducing the multivariate shifted exponential normal (MSEN) distribution, an elliptical heavy-tailed generalization of the multivariate normal (MN). The MSEN belongs to the family of MN scale mixtures (MNSMs) by choosing a convenient shifted exponential as mixing distribution. The probability density function of the MSEN has a closed-form characterized by only one additional tailedness parameter, with respect to the nested MN, governing the tail weight. The first four moments exist, and the excess kurtosis can assume any positive value. The membership to the family of MNSMs simplifies maximum likelihood (ML) estimation of the parameters via the expectation-maximization (EM) algorithm; advantageously, the M-step is computationally simplified by closed-form updates of all the parameters. Since the tailedness parameter is estimated from the data, robust estimates of the mean vector of the nested MN distribution are automatically obtained by down weighting; we show this aspect theoretically but also by means of a simulation study. We fit the MSEN distribution to multivariate allometric data where we show its usefulness also in comparison with other well-established multivariate elliptical distributions.

E0876: **Enriching the AR(p) process with skewed generalised normal innovations**
Presenter: **JT Ferreira**, University of Pretoria, South Africa
Co-authors: Ane Neethling, Andriette Bekker, Mehrdad Naderi

The autoregressive (AR) process of order p is of common value in many areas of research, not necessarily only in a time series environment. Usually, innovations are assumed to be iid normally distributed; this assumption may not characterise some true processes adequately due to the normal distributions restrictions regarding asymmetry. A skew generalised normal distribution is presented as an alternative to the usual normal assumption. It can account for not only skewness but also heavier tails that innovations might exhibit. This model is compared to previously proposed models, and its position as a valid contender as the choice of innovation process discussed via some simulation studies and data analysis.

E0993: **Bayesian analysis of skew spherical data with rotationally-symmetric distributions**
Presenter: **Najmeh Nakhaeirad**, University of Pretoria, South Africa
Co-authors: Andriette Bekker, Mohammad Arashi

The von-Mises Fisher distribution is the most common symmetric distribution in spherical studies. In reality, the symmetry assumption of the underlying distribution is often rejected. This shows the importance of studying the skewed directional distributions. Therefore, the skewed version of von-Mises Fisher distribution is a good alternative. Bayesian inference of skew-von-Mises Fisher distribution is presented. Sample generation from the posterior distribution is discussed using the modified Gibb's sampling and the weighted bootstrap resampling algorithms and to illustrate the obtained results a numerical example is provided.

EO325 Room R18 MODEL VALIDATION Chair: Maria Dolores Jimenez-Gamero

E0577: **On Stein operators in testing the fit to parametric families of distributions**
Presenter: **Bruno Ebner**, Karlsruhe Institute of Technology, Germany

Characterisations of families of distribution are widely used as a basis for proposing tests of fit to parametric families of distributions. These families include classical problems like testing normality, exponentiality or Poissonity. We review the utility of Stein operators as well as Stein type characterising identities in goodness-of-fit testing and propose new ways to approach the problem. We show that these classes of tests are theoretically well understood and present themselves as serious competitors to other existing tests.

E0994: **Thematic-accuracy quality control of slope and aspect classes based on an equivalence test**
Presenter: **Virtudes Alba-Fernandez**, University of Jaen, Spain
Co-authors: Francisco Javier Ariza-Lopez, Maria Dolores Jimenez-Gamero

The thematic quality control of slope and aspect classes derived from a digital elevation model can be performed comparing the results on a product and a reference using a confusion matrix. Because of the nature of data (elevations and derived slopes and aspects), users can be more interested in checking proximity than equality toward the reference quality requirements. In the spirit of incorporating little or irrelevant deviations between the observed data and the target population in the definition of the hypothesis, a model equivalence test is proposed. Such deviations are assessed utilizing a ϕ-divergence measure, which in turn can be consistently estimated. The asymptotic behavior of the resulting test statistic is studied, and a critical region based on the asymptotic null distribution is considered. The finite sample performance of the proposal has been studied through several simulation experiments. Finally, the new test is conducted over a real data set.

E0991: **New distribution-free goodness-of-fit tests for the Pareto distribution**
Presenter: **Bojana Milosevic**, University of Belgrade, Serbia

Co-authors: James Allison, Marko Obradovic, Lizanne Raubenheimer, Marius Smuts

Three new classes of goodness-of-fit tests for the Pareto type I distribution are proposed based on an equidistribution type characterization. The asymptotic null distributions are derived, and the Bahadur efficiencies of the new tests are compared to the efficiencies of some existing tests. The finite-sample behaviour is also studied through extensive power comparison. It is found that the integral type test performs the best among the new tests and that it also performs favourably in comparison to competitor tests.

E1032: **A class of goodness-of-fit tests for the Rayleigh distribution based on conditional expectation**

Presenter: **James Allison**, Northwest University, South Africa

Co-authors: Joseph Ngatchou-Wandji

New goodness-of-fit tests for the Rayleigh distribution are proposed and studied. The tests are based on a characterization involving a conditional expectation. The asymptotic properties of the tests are explored. The performance is evaluated and compared to that of existing tests by means of a Monte Carlo study. It is found that the newly proposed tests perform satisfactory compared to the competitor tests.

E0974: **Estimating the shape functions**

Presenter: **Maria Dolores Jimenez-Gamero**, Universidad de Sevilla, Spain

The left shape function and the right shape function have several remarkable properties. Among them, if F is a location-scale family of continuous distribution functions, then any random variables X and Y with distribution function in this family have the same right (left) shape function. In other words, the right (left) shape function characterizes location-scale families. This property can be used to built goodness-of-fit tests. A key step towards the development of statistical procedures based on the right shape function for making inferences is the study of an estimator of such function. This is the objective.

EO087 Room R21 ADVANCES IN SURVIVAL AND RELIABILITY I Chair: Juan Eloy Ruiz-Castro

E0788: **Estimation of a model for spatial binary data**

Presenter: **Gabrielle Kelly**, University College Dublin, Ireland

A model for a realization of a binary random field is considered where the correlations satisfy the Frechet-Hoeffding bounds. The binary variables are related to latent variables that have a Matern spatial correlation. Thus, both the marginal means of the binary variables and their spatial distances contribute to their spatial correlations. The model is fitted to TB infection data in cattle herds and wildlife badgers with a fixed censoring time of 2 years. Estimates of the practical range are obtained, i.e. distances at which spatial correlation can be regarded as negligible.

E0903: **Estimation of bivariate survival functions: A simulation study on the effects of sample size**

Presenter: **Marialuisa Restaino**, University of Salerno, Italy

Bivariate survival data have received considerable attention recently. In survival analysis, it is common to deal with incomplete information of the data, due to random censoring and random truncation. Most of the existing research focuses on bivariate survival analysis when components are either censoring or truncation or when one component is censored and truncated, but the other one is fully observed. Moreover, due to missing information related to censoring and truncation, it becomes crucial to have an adequate sample size, to have a significant estimate of the bivariate survival function. Starting from this background, after reviewing the most used estimators for the bivariate survival function, when both components are censored and truncated, we will inspect the effects of different sample sizes of the bivariate survival functions, according to some censoring percentages and truncation probabilities. By a simulation study and application to real datasets, we will test the influence of sample sizes on the performance of the estimators.

E0956: **Multistate model for trajectories clustering**

Presenter: **Rossella Miglio**, Bologna University, Italy

Clustering of temporal or sequential data is challenging, especially when dealing with discrete data. The motivating problem is to find patterns of drug use trajectories over time. It is essential to have standard measures of change, define appropriate similarities among trajectory observations, obtain appropriate data representation and use methods that are suitable for this kind of data or using information extracted from them to apply classical methods. We analysed data across 5 years on a sample of 70000 drug users to identify transition patterns in drug use trajectories during this time. Data were collected every three months for a total of 20 measurements for each subject, demographic and some clinical covariates are available. Optimal matching and a three-step procedure proposed previously to identify clusters of individual longitudinal trajectories were used in the preliminary analysis. We propose to address the problem of unsupervised classification of sequences using a multi-state approach, to obtain measures to quantify the change in drug use behaviours; these approaches allow us to consider also the effects of covariates. These set of measures could be used as input for classical clustering methods but could also help to provide effective visualizations for applied use. The results obtained by the other proposed method are compared with this new proposal.

E1034: **Analysing the Lombardy regions geriatric wards**

Presenter: **Hannah Mitchell**, Queen's University Belfast, United Kingdom

Co-authors: Hannah Mitchell, Adele Marshall, Mariangela Zenga

Modelling patient flow within a healthcare setting is seen as an important aspect of understanding the activity of the system. Healthcare managers are under increased pressure, particularly during these uncertain times, to maintain a high quality of care within already strained healthcare systems. Previous research into the Coxian phase-type distribution has demonstrated its effectiveness in modelling patient flow through the hospital system. By joining the Coxian phase-type distribution with the continuous-time hidden Markov model (Coxian-CTHMM) enables different inherent pathways to be identified, which in turn provides healthcare managers with a deeper understanding of the healthcare system. The Coxian phase-type distribution and the Coxian-CTHMM will be applied to administrative data for the Lombardy regions geriatric wards in 2015.

E1059: **Deep learning for time to event data**

Presenter: **Federico Ambrogi**, University of Milan, Italy

Co-authors: Thomas Scheike

The use of neural networks for developing prediction models with survival data has received much attention in the literature. The developments with machine learning methods and with computational facilities have renewed interest in such kind of models and with the use of deep neural networks. There are some proposals already available for adapting deep learning regression models to time to event data. We propose a simple method, easily generalisable to complex settings, such as competing risks or semi-competing risks, working with standard software for deep neural networks in R and Python, namely Keras. The presented approach is based on the use of standard binomial estimating equations with weights. A similar approach is the one based on pseudo values allowing to use standard software provided it is possible to have an appropriate link function. An application based on SEER data is presented with a comparison of model calibration and prediction error with respect to standard methods.

EC786 Room R08 CONTRIBUTIONS IN COMPUTATIONAL STATISTICS Chair: Matthieu Marbac

E1088: **Using full ranking information in ranked set and judgment post-stratification sampling designs**

Presenter: **Mahdi Salehi**, University of Pretoria, South Africa

Co-authors: Bardia PanahBehagh, Mohammad Salehi

Ranked set sampling (RSS) and judgment post-stratification sampling (JPS) are two efficient sampling designs where one can rank units without full

measurement based on visual inspection or some auxiliary variables. In all of the different available variants of RSS and JPS, many sets are selected to rank the units, but the information of just one set will be used for ranking each unit. Even for those versions of RSS using multi-ranker, one utilizes only the information from one set. In contrast to all the previous methods, we propose two approaches for improving the existing methods of RSS and JPS by employing all the selected sets to gain more information about the rank of units to be measured. Indeed, the information of all the sets will be used for all the selected units to be measured; consequently, any improper information of some potential outlier sets can be adjusted by the other sets. The mean estimators are developed. Using a relatively comprehensive simulation study, some other well-known ranked based competitors, including the median and the extreme RSS plans with their new versions constructed based on the introduced designs are compared as well. The results show that new designs lead to more efficient estimators than the ordinary counterparts' estimators for all considered cases.

E1138: **Multilevel bootstrap particle filter**
Presenter: **Daniel Burrows**, University of Bath, United Kingdom
Multilevel Monte Carlo (MLMC) was introduced at the turn of the millennium, and it has since become popular, particularly in the field of financial mathematics as a means to avoid expensive solving of stochastic differential equations. We develop MLMC in the context of sequential Monte Carlo (SMC) to obtain a novel multilevel bootstrap particle filter (MLBPF). SMC methods are often the only feasible way to estimate the distribution of a partially observed latent Markov process. We consider situations where the evaluation of the particle weights, i.e. the likelihood, is too expensive to make the standard bootstrap particle filter (BPF) feasible for large sample sizes. We adopt the multilevel approach whereby a substantially less expensive likelihood approximation is used for the bulk of particles, and the resulting bias due to the approximation is then corrected by a relatively small number of expensive evaluations of the exact likelihood. As a result, we obtain an approximation accuracy comparable to BPF, but at a lower computational cost. We establish the strong law of large numbers and central limit theorem for MLBPF and demonstrate its performance on numerical applications.

E1151: **Inference from non-probability surveys with XGBoost**
Presenter: **Luis Castro Martin**, University of Granada, Spain
Co-authors: Ramon Ferri-Garcia, Antonio Arcos
The importance of online methods for data collection has increased the relevance of non-probability surveys, since they depend on self-selection procedures and suffer from coverage problems. These issues result in biased estimations. Some techniques like Statistical Matching and Propensity Score Adjustment have been proposed to compensate this bias using an auxiliary probability sample. However, they usually rely on logistic or linear regression as the chosen machine learning algorithm. We propose using a state-of-the-art algorithm like XGBoost instead. Simulation studies are conducted to demonstrate how much can XGBoost improve the current results.

E0629: **p-variation of cusum process and testing change in the mean**
Presenter: **Tadas Danielius**, Vilnius University, Lithuania
A new test of model instability in the mean is proposed and investigated. The test is based on p-variation of step-wise cusum process. We establish a limiting distribution of the test statistics under null as well under contiguous alternatives.

E0317: **A scaled LASSO for multicollinear situations**
Presenter: **Mohammad Arashi**, Ferdowsi University of Mashhad, Iran
Co-authors: Yasin Asar, Bahadir Yuzbasi
A scaled LASSO is proposed by pre multiplying the LASSO with a matrix term, namely scaled LASSO (SLASSO) for multicollinear situations. A numerical study shows that the SLASSO is comparable with other thresholding techniques and often outperforms the LASSO and elastic net.

EC449 Room R11 CONTRIBUTIONS IN MULTIVARIATE STATISTICS Chair: Christian Hennig

E0990: **Non-Gaussian component analysis: Testing the dimension of the signal subspace**
Presenter: **Klaus Nordhausen**, Vienna University of Technology, Austria
Co-authors: Una Radojicic
Dimension reduction is a common strategy in multivariate data analysis which seeks a subspace which contains all interesting features needed for the subsequent analysis. Non-Gaussian component analysis attempts for this purpose to divide the data into a non-Gaussian part, the signal, and a Gaussian part, the noise. We will show that the simultaneous use of two scatter functionals can be used for this purpose and suggest a bootstrap test to test the dimension of the non-Gaussian subspace. Sequential application of the test can then, for example, be used to estimate the signal dimension.

E1131: **A unite-and-conquer-based approach that improves weak classification results**
Presenter: **Abdoulaye Diop**, University of Versailles Paris Saclay, France
Co-authors: Nahid Emad, Thierry Winter
In the field of machine learning, statistical classification methods are used as a solution to multiple problems. The main idea of these methods is to build discriminative models that create decision boundaries that separate the classes. However, depending on the properties of the data studied, these models may exhibit poor class prediction performance. This problem can be a consequence of statistical issues such as class imbalances, high bias, and high variance. We propose a framework based on ensemble learning techniques and a unite and conquer approach to deal effectively with these problems. This approach makes it possible to manage the bias-variance trade-off and improve the training time and the results of the base methods composing the ensemble learner. With the detection of behavioral anomalies as a case study, we show the interest of this approach for its improvement of the prediction results and its efficiency on high-performance computing systems.

E0835: **Sparse PLS-DA: Clustering time series for art conservation**
Presenter: **Sandra Ramirez**, Universidad Politecnica de Valencia, Spain
Co-authors: Manuel Zarzo, Fernando-Juan Garcia-Diego, Angel Perles
Clustering time series data has a wide range of applications. One problem when analyzing time series for art conservation is that time series of relative humidity RH are too similar, even when positioned differently. Many studies have displayed that one common problem is that if underlying clusters are very close to each other, the clustering performance might diminish significantly. Before applying the discriminant technique, the variables that are extracted from the time series were determined. They correspond to estimates of parameters from time series models and features from time series functions. The number of variables was greater than the number of time series. The goal is to propose a methodology for classifying time series when they are similar and have more variables. Sparse Partial Least Squares Discriminant Analysis (sPLS-DA) was applied using three prediction distances and two classification error rates. The methodology is described using different time series of RH from a study carried out in the Metropolitan Cathedral of Valencia in 2008 and 2010. This methodology would be perfect for applying to time series data of relative humidity from both churches and museums, or similar buildings where it is possible to indicate the class of the time series.

E0807: **Simultaneous feature selection and outlier detection with optimality guarantees**
Presenter: **Luca Insolia**, Scuola Normale Superiore, Italy
Co-authors: Ana Kenney, Francesca Chiaromonte, Giovanni Felici
Sparse estimation in the presence of outliers has received considerable attention in the last decade. We contribute by considering high-dimensional

regression models contaminated by multiple mean-shift outliers affecting both the response and the design matrix. We develop a general framework and use mixed-integer programming to simultaneously perform feature selection and outlier detection with provably optimal guarantees. We prove theoretical properties for our approach, i.e., a necessary and sufficient condition for the robustly strong oracle property, where the number of features can increase exponentially with the sample size; the optimal estimation of parameters; and the breakdown point of the resulting estimates. Notably, our proposal requires weaker assumptions than prior methods in the literature and, unlike such methods, it allows the sparsity level and/or the amount of contamination to grow with the number of predictors and/or the sample size. Moreover, we provide computationally efficient procedures to tune integer constraints and warm-start the solution algorithm, and, through simulations, show the superior performance of our proposal with respect to existing heuristic methods. Finally, the method is deployed to elicit the role of microbiome in childhood obesity.

E1062: **Relative complexity assessment in multifractal systems: Application to spatio-temporal seismic dynamics**
Presenter: **Francisco Javier Esquivel**, University of Granada, Spain
Co-authors: Francisco Javier Alonso, Jose Miguel Angulo

The structural complexity of phenomena with multifractal dynamics and possible disequilibrium states has been studied, in particular, using information-theoretic approaches. A direct relationship is established between the scaling behaviour of Renyi-entropy-based generalized product complexity measures and the increments of generalized Renyi dimensions. Beyond their marginal assessment, the degree of local coherence between interacting magnitudes in a system can inform about eventual structural changes in their joint spatio-temporal dynamics. From a divergence-based formulation of generalized relative Renyi dimensions, an analogous limiting relation with Renyi-divergence-based generalized product relative complexity measures is proved. In both cases, a significant interpretation is given in terms of sensitivity of marginal and relative diversity indices with respect to the deformation parameter. A spatio-temporal analysis of seismic data is performed using these measures, focusing on the detection of structural changes in the association between the frequency of events and the accumulated released energy distributions.

EC796 Room R22 CONTRIBUTIONS IN BAYESIAN STATISTICS Chair: Bernardo Nipoti

E0892: **Ensemble sampler for infinite-dimensional inverse problems**
Presenter: **Jeremie Coullon**, Lancaster University, United Kingdom
Co-authors: Robert Webber

A Markov chain Monte Carlo (MCMC) sampler for infinite-dimensional inverse problems is introduced. The proposed MCMC sampler is more efficient than preconditioned Crank-Nicolson, yet it is easy to implement and it requires no gradient information or posterior covariance information. The new sampler involves truncating the Karhunen-Loeve expansion of the function to decompose the space into a low-dimensional subspace that is mainly influenced by the likelihood, and the complementary subspace which approximately follows the prior distribution. The affine invariant ensemble sampler is used for proposals in the low-dimensional space while pCN is used for the complementary subspace. As a result, the new sampler is able to handle strongly correlated posteriors, is self-tuning, and is easy to parallelise. We use two numerical experiments to compare the performance of the new sampler to pCN and an existing gradient-free sampler in the literature. The first example involves inferring the wave speed and initial condition of the advection equation, and the second example involves parameter estimation for Langevin dynamics where we infer the two scalar parameters along with the posterior position path X_t. We find that the new sampler is more robust and outperforms both methods. We conclude with a discussion of the limitations of the method and avenues for future research.

E0963: **Bayesian nonparametrics for sparse dynamic networks**
Presenter: **Cian Naik**, University of Oxford, United Kingdom
Co-authors: Francois Caron, Judith Rousseau, Yee Whye Teh, Konstantina Palla

A Bayesian nonparametric approach for sparse time-varying networks is proposed. A positive parameter is associated with each node of a network, which models the sociability of that node. Sociabilities are assumed to evolve over time and are modelled via a dynamic point process model. The model is able to (a) capture long term evolution of the sociabilities, and (b) yields sparse graphs, where the number of edges grows subquadratically with the number of nodes. The evolution of the sociabilities is described by a tractable time-varying generalised gamma process. We provide some theoretical insights into the model and apply it to three real-world datasets: a network of hyperlinks between communities on Reddit, email exchanges between members of the Democratic National Congress, and co-occurrences of words in Reuters news articles after the September 11^{th} attacks.

E1136: **Challenges and proposals for Dirichlet process mixture models (DPMM) with Gaussian kernels**
Presenter: **Wei Jing**, University of St Andrews, United Kingdom
Co-authors: Michail Papathomas, Silvia Liverani

The Dirichlet process mixture model (DPMM) is considered in the context of clustering for continuous data when the conditional likelihood is set to be the multivariate normal distribution. Simulation studies show that the DPMM struggles to uncover true clusters when the data contain even just a handful of variables, even when the normality assumption is correct. An introduction of the DPMM is given first, followed by simulation examples highlighting the problem the DPMM currently faces. Potential reasons that lead to the problem are analyzed. Specifically, one of the reasons is the difference between the overall covariance matrix for the variables (calculated from pooling the data of all the clusters) and the within-cluster covariance matrices, which impedes the sampler from moving towards the target cluster allocation. Another possible factor is adopting an unsuitable prior distribution for the within-cluster covariance matrices. Different priors that can be placed on the within-cluster covariance matrices are reviewed, and their performance is assessed and compared. Finally, other aspects that can improve or influence the performance of the DPMM are discussed.

E0926: **Bayesian approach for uplift linear regression**
Presenter: **Yuji Iikubo**, Waseda University, Japan
Co-authors: Shunsuke Horii, Toshiyasu Matsushima

Uplift modeling is an important method for predicting the difference in effect when an action is taken and when it is not taken. Previously, many studies for uplift modeling have focused on the classification problem. However, their performances have been evaluated experimentally due to the difficulty of theoretical analysis. On the other hand, a few studies have focused on regression problems for theoretical analysis. The proposed estimators are unbiased and asymptotically normal, and they have focused on reducing the variance of the estimators. We propose a Bayesian approach for uplift linear regression. Assuming multivariate normal distributions for the prior distributions of the parameters, we show that the Bayes optimal estimator and predictor can be obtained in closed forms. In uplift modeling problems, another important issue is to determine the assignments of actions to estimate efficiently with small amounts of data. One of the advantages of the Bayesian approach is that you can obtain the posterior distributions of the parameter and the predictor. We also propose a sequential experimental design method by taking advantage of the Bayesian approach. Finally, we show the effectiveness of the proposed methods by numerical experiments using synthetic data and semi-synthetic data.

E0869: **Using Bayesian change-point Markov sampler in basecalling of nanopore signals**
Presenter: **Sophia Shen**, Macquarie University, Australia
Co-authors: Georgy Sofronov

DNA sequencing is an important subdiscipline in bioinformatics for there are many crucial applications, such as in forensic to convict criminals or in

medicine to diagnose diseases like cancer or devise personalised drug interventions. The latest DNA sequencing technologies using enzyme-based nanopores are capable of capturing long repetitive DNA structures frequently present in introns. The commercialisation of the portable nanopore sequencer MinION made DNA sequencing more accessible despite its high error rates. The process of translating raw (electrical) nanopore signals into genetic alphabets is called basecalling. Statistical methods such as hidden Markov models (HMM) and recurrent neural networks (RNN) have been employed to analyse DNA bases during basecalling. An alternative algorithm is considered based on the Gibbs sampler that allows transitions between different models. The change point framework is adopted as each base transition can be thought of as a change point in the nanopore signals. Our numerical study shows that the proposed Bayesian change point algorithm can identify the number of change points and their locations at the same time and its flexibility have the potential to make DNA base identification more robust during basecalling.

EG084 Room R15 CONTRIBUTIONS IN CAUSAL INFERENCE AND GRAPHICAL MODELS Chair: Tetiana Gorbach

E0775: **Bayes optimal estimator of the mean intervention effect and its approximation based on variational inference**
Presenter: **Shunsuke Horii**, Waseda University, Japan

To estimate the causal effect under Structural Causal Models (SCMs), it has to know or estimate the model that generates the data. However, it is often difficult or impossible to verify which model is correct based only on the data, so some models remain as candidates for the data generating model. We first show from a Bayesian perspective that it is Bayes optimal to weight (average) the causal effects estimated under each model rather than estimating the causal effect under a fixed single model. This idea is also known as Bayesian model averaging, and we attempt to apply it to the causal estimation under SCM. Although the Bayesian model averaging is optimal, as the number of candidate models increases, the weighting calculations become computationally hard. We develop an approximation to the Bayes optimal estimator by using the variational Bayes method. We show the effectiveness of the proposed methods through numerical experiments based on synthetic and semi-synthetic data.

E1002: **Graphical modelling of multivariate panel data models**
Presenter: **Celia Gil-Bermejo**, Universidad Complutense de Madrid, Spain
Co-authors: Antonio Jesus Sanchez Fuentes, Jorge Onrubia

A new approach is proposed to both test the existence of causal relationships between variables in a panel data environment using a VAR model and determine one final causality path excluding those relationships which are redundant. One of the main novelties is that we extend the number of relevant variables, mostly limited to two/three variables when using panel data. Once we set the dependence criteria (in our case, the concept of Granger causality), we apply the PC algorithm in order to debug potential indirect relationships between variables. This algorithm uses an iterative process where different conditional tests between each pair of variables are carried out. Thanks to these individual measures, we construct one synthetic measure for the whole sample. Finally, once the causality path between all the possible combinations of variables has been established, we draw it using causal maps. These figures provide a visual guide which makes explicit complex interlinked relationships. Moreover, this approach helps us to analyse the determining factors influencing these relationships.

E0584: **Variational inference for sparse high-dimensional graphical-VAR models**
Presenter: **Nicolas Bianco**, University of Padua, Italy
Co-authors: Mauro Bernardi, Daniele Bianchi

The increased complexity of modern datasets requires suitable techniques to select the most relevant features and to carry out an accurate inference in a reasonable amount of time. We develop a variational approximation algorithm to deal with sparse estimation of high-dimensional graphical vector autoregressive models with the possibility of including some exogenous covariates. The purpose is two-fold. First, we exploit the product density factorisation of the joint variational density that leads to the mean-field paradigm as well as the representation of the problem as a sequence of auxiliary regressions that rely on the Cholesky factorisation of the precision matrix. Both the Normal-double-Gamma prior and the Spike and Slab prior are implemented to shrink toward zero both the autoregressive and the precision matrices. The second contribution concerns the solution of the lack-of-identification problem that relies on the employed Cholesky factorisation. We propose to approximate the marginal likelihood of each model permutation by the variational model evidence and to exploit it to get the MaP estimates of the model parameters. When the dimension of the model is large, the complete exploration of the permutations' space becomes unfeasible, hence a parallel interacting simulated annealing algorithm is used in this case.

E0983: **Uplift modeling with multitreatment for observational pretest-posttest data**
Presenter: **Hiroaki Naito**, Doshisha University, Japan
Co-authors: Hisayuki Hara

Uplift modeling is used to optimize the effect of intervention by predicting the causal effect for each unit from its covariates. Uplift refers to the causal effect for each value of covariates. Uplift modeling has been developed in a randomized controlled trial like A/B test of direct marketing. Recently, uplift modeling is also extended to the nonrandomized study, where only observational data are available. For observational data, a switch doubly robust method (SDRM) has been proposed recently. SDRM is based on doubly robust (DR) estimator and avoids instability of DR estimator due to extreme propensity scores. SDRM considers the case where the treatment assignment is binary (whether or not treatment has been received). Currently, SDRM is the state of the art of uplift modeling for observational cross-sectional data. We will extend SDRM for observational pretest-posttest data. Pretest-posttest data consists of outcomes before and after the intervention and are often used to evaluate causal effects qualitatively. In addition, we will extend the treatment assignment to multi-treatment. This extension allows for more complex intervention strategies. We will perform some simulation studies and apply the proposed method to real data example to confirm the usefulness of the proposed methods.

E0879: **jewel: A novel method for joint node-wise estimation of multiple Gaussian graphical models**
Presenter: **Anna Plaksienko**, Gran Sasso Science Institute, Italy
Co-authors: Claudia Angelini, Daniela De Canditiis

Graphical models are well-known mathematical objects for describing conditional dependency relationships between random variables of a complex system. Gaussian graphical models refer to the case of multivariate Gaussian variable for which the graphical model is encoded through the support of corresponding inverse covariance (precision) matrix. We consider a problem of estimating multiple Gaussian graphical models from high-dimensional data sets under the assumption that they share the same conditional independence structure. However, the individual correlation matrices can differ. Such a problem can be motivated by applications where data comes from different sources and can be collected in distinct classes or groups. We propose a joint data estimation that uses a node-wise penalized regression approach. Grouped Lasso penalty simultaneously guarantees the resulting adjacency matrix's symmetry and the joint learning of the graphs. We solve the minimization problem using the group descent algorithm and establish the proposed solution's consistency and sparsity properties. Finally, we show how the regularization parameter can be estimated using cross-validation and BIC. We provide a novel R package jewel with the implementation of the proposed method and illustrate our estimator's performance through simulated and real data examples. We compare the proposed approach with other available alternatives.

EG074 Room R17 CONTRIBUTIONS IN TIME SERIES Chair: Maddalena Cavicchioli

E0225: **Testing structural breaks: A new self-normalization approach based on the adjusted sample range**
Presenter: **Jiajing Sun**, University of Chinese Academy of Sciences, China
Co-authors: Yongmiao Hong, Brendan McCabe

The aim is to test for structural breaks, and to propose a new self-normalization approach based on the adjusted range of the partial sum process. First, we extend the Kolmogorov-Smirnov test statistic using the adjusted range based self-normalization to test for a change in the mean. Second, we extend previous and introduce the G statistic, based on the adjusted range of the partial sum, which can accommodate multiple structural changes and structural changes in a multi-dimensional setting. Third, we extend the range-normalized KS and G statistics to consider structural changes in a general setting. Fourth, we extend the range-normalized KS and G test statistics to testing parameter constancy under a conditional autoregressive (CAR) framework. The focus on CAR follows from the superiority of the range-based self-normalization under conditions of persistent autocorrelation. However, parameter constancy tests developed can be easily generalized to other tests of parameter constancy under suitable conditions. We explore the statistical properties of these test statistics and conduct simulation and empirical studies. Our results show that the range-based test statistics are amongst the most reliable offering monotonic power when the nave self-normalization and kernel based long-run variance estimators fail to do so. The empirical studies also reveal that range-based test statistics are of great help in testing for structural breaks.

E0238: **High-dimensional sparse multivariate stochastic volatility models**
Presenter: **Benjamin Poignard**, Osaka University, Japan
Co-authors: Manabu Asai

Although multivariate stochastic volatility models usually produce more accurate forecasts compared to multivariate GARCH models, their estimation techniques such as Bayesian Markov Chain Monte Carlo are computationally demanding and thus suffer from the curse of dimensionality. We propose a fast estimation approach for MSV models based on a penalised ordinary least squares framework. We propose a two-step penalised procedure for estimating the latter using a broad range of potentially non-convex penalty functions. This two-step procedure relies on OLS based loss functions and thus easily accommodates high-dimensional vectors. We provide the large sample properties of the two-step estimator together with the oracle property of the first step estimator with a diverging number of parameters. The empirical performances of our method are illustrated through in-sample simulations and out-of-sample forecasts.

E0665: **Segmentation of autoregressive processes via the cross-entropy method**
Presenter: **Lijing Ma**, Macquarie University, Australia
Co-authors: Georgy Sofronov, David Bulger

One approach to modelling nonstationary time series data is to use change point detection methods to optimally segment the signal into intervals within which the process behaves stationarily. We assume that each segment is an autoregressive time series with its own model parameters. Taking the nonstationarity into account and identifying the number and locations of these structural breaks are of interest. Our method includes two steps. The first step is to use a distributionally tailored cross-entropy method to identify these potential change points to segment the time series. Once these potential change points are obtained, modified parametric spectral discrimination tests are used to validate the proposed segments. A numerical study is conducted to demonstrate the performance of the proposed method across various scenarios and compare it against other contemporary techniques.

E0893: **Spectral analysis of Markov switching GARCH models**
Presenter: **Maddalena Cavicchioli**, University of Modena and Reggio Emilia - Dipartimento di Economia Marco Biagi, Italy

Matrix expressions in closed form are derived for the autocovariance function and the spectral density of Markov switching GARCH models and their powers. For this, we apply the Riesz-Fischer theorem which defines the spectral representation as to the Fourier transform of the autocovariance function. Under suitable assumptions, we prove that the sample estimator of the spectral density is consistent and asymptotically normally distributed. Further statistical implications in terms of order identification and parameter estimation are discussed. These methods are well suited for financial market applications, and in particular for the analysis of time series in the frequency domain, as shown in the proposed numerical and real-world examples.

E1189: **Nonlinear causality for CHARN models**
Presenter: **Xiaoling Dou**, Waseda University, Japan

The CHARN model was proposed in financial data analysis. Because of its non-normality, non-linearity and the blindingly general form, it has come into use in various fields of time series. We consider a nonlinear causality for the CHARN models. We show that the causality of the CHARN models can be evaluated by a Portmanteau test, based on a constrained maximum likelihood estimator of the parameters, and the test statistic has an approximate asymptotic Chi-square distribution. We describe the Chi-square Asymptotics of the Portmanteau test for a CHARN model, provide calculations of the test statistic and investigate the performance of the Portmanteau test by simulation.

EG010 Room R20 CONTRIBUTIONS IN APPLIED STATISTICS I Chair: Georg Lindgren

E0589: **Policing route optimization via density-based principal curves**
Presenter: **Ben Moews**, The University of Edinburgh, United Kingdom
Co-authors: Jaime R Argueta, Antonia Gieschen

A new method is introduced for identifying patrol routes in hot spots through ridge estimation to explore the application of density ridges to hot spots and patrol optimization, and to contribute to the literature in police patrolling. We make use of the subspace-constrained mean shift algorithm, a recently introduced approach for ridge estimation further developed in cosmology, which we modify and extend for geospatial datasets and hot spot analysis. The experiments extract density ridges of Part I crime incidents from the City of Chicago during the year 2018 and early 2019, with results demonstrating nonlinear mode-following ridges in agreement with broader kernel density estimates. Using early 2019 incidents with predictive ridges extracted from 2018 data, we create multi-run confidence intervals and demonstrate near-complete coverage with narrow envelopes around ridges. We also develop and provide researchers, as well as practitioners, with a user-friendly and open-source software for fast geospatial density ridge estimation. Our empirical tests show the stability of ridges based on past data, offering an accessible way of identifying routes within hot spots instead of patrolling epicenters. We suggest further research into the application and efficacy of density ridges for patrolling.

E0699: **Statistical methods for space surveillance**
Presenter: **Antonio Arcos**, Universidad de Granada, Spain

Since the first human spacecraft, Sputnik-1, the Earth proximities have been occupied by multiple human-made objects which have created a particle environment known as "Space Debris". The mitigation of this phenomena is a crucial safety task nowadays, therefore adopting measures to relieve and prevent these threats has aroused a worldwide concern. Modern space systems require the highest possible accuracy to work efficiently. Sensors measurements should be differentially corrected to actually determine de real and precise orbit in what is called Statistical Orbit Determination (SOD). The application of predicting filters, as Kalman or particle, allows the system to rapidly and memoryless solve nonlinear least-square problems simultaneously estimating not only the state but also the covariance matrix of the target. Besides the SOD part, other statistical methods are applied to model the different parts of the tracker architecture, solving likelihood optimal associations with several algorithms. A simulated scenario is created to model various space environments and its measurements to conclude in an optimal configuration of the tracker; in which it could be remarked the performance of the Unscented Kalman Filter (UKF) and the Joint Probabilistic Data Association (JPDA) algorithm.

E0875: **A regularized model for spectroscopic data**
Presenter: **Chin Gi Soh**, Nanyang Technological University (National Institute of Education), Singapore
Co-authors: Ying Zhu

High-dimensional spectroscopic data is informative, and has applications in many fields such as biomedical sciences and food science. The fitting of regression models for the purposes of prediction is known to be a challenging task due to the high dimension of the datasets, as well as the high correlation between wavenumbers in the data. One method that has gained interest in recent years is the use of regularization to overcome these challenges. We present a regularized model for spectroscopic data. The penalty functions used are designed to capture the underlying group structure in the spectroscopic data, as well as to give rise to an interpretable model. Depending on the penalty functions used in the regularized model, different computational challenges may arise. We will discuss some algorithms that are of interest in solving for such regularized model coefficients, as well as the advantages and disadvantages of these algorithms. An example of the application of the model to Fourier-transform infrared spectroscopic data for the prediction of olive oil purity will be presented.

E1172: **Statistical modelling of mechanical skin behaviour**
Presenter: **Maria Filomena Teodoro**, CINAV - Portuguese Naval Academy, Portugal
Co-authors: Teresa Oliveira

Several studies have been carried out on the gender difference, in the most different variables. Between men and women, there are biological and physiological differences, which are revealed at the behavioural and emotional level. The skin and adjacent soft tissues of each individual assume a mechanical behaviour when subjected to external forces. Exploring the distinct behaviour of male and female, the aim is to investigate the differences between genders concerning the perception of pain and soft tissue deformation, complementing, with the evaluation of BMI and fat fold. Male and female individuals were tested in several anatomical regions. An evaluation of the psychological behaviour and physical condition of individuals was performed through questionnaires, before and after the application of the indentation test. All individuals were subject of several measurements and the indentation test (Strengths, Deformations and absorbed Energy). We describe the results obtained by analysis of variance approach. Possible associations between the different measured variables were investigated. It was concluded that there are significant differences between genders, both in the measured variables and in the variables related to the indentation test. Given these differences, the variable maximum strength stands out, which presents a significant difference between the maximum value for each gender.

E0995: **Analysis of aggregate zonal imbalance in the Italian electricity market using copula models**
Presenter: **Aurora Gatto**, University of Salento, Italy
Co-authors: Fabrizio Durante, Francesco Ravazzolo

The problem of determining possible correlations and dependencies with the aggregate zonal imbalance in electricity markets is considered. In particular, we are interested in the estimation of a model for the aggregate zonal imbalance, that is, the algebraic sum, changed in sign, of the amount of energy procured by the Italian national Transmission and System Operator (Terna) in the Dispatching Services Market (Msd) at a given time in a given Italian electricity macro-zone. From a methodological point of view, we use a model that combines the analysis of classical time series with copula-type models. As a result, the flexibility of a copula approach will allow identifying the nature of non-linear linkage among the aggregate zonal imbalance and other variables of interest for the electricity market such as forecasted demand, forecasted wind and solar PV generation.

CI021 Room R04 ADVANCES IN BAYESIAN ANALYSIS AND APPLICATIONS Chair: Mike So

C0394: **Clustering large scale generalized linear longitudinal models with grouped patterns of unobserved heterogeneity**
Presenter: **Tomohiro Ando**, Melbourne Business School, Australia
Co-authors: Jushan Bai

Methods are provided to flexibly capture the unobservable heterogeneity from longitudinal data in the context of the exponential family of distributions. The group membership of individual units is left unspecified, and their heterogeneity is influenced by group-specific unobservables as well as the heterogeneous regression coefficients. We discuss a computationally efficient estimation method and derive asymptotic theory. The established asymptotic theory includes a uniform consistency of the estimated group membership. To test the heterogeneous regression coefficients within-group or not, we propose the Swamy-type test that takes account unobserved heterogeneity. We apply the proposed method to study the market structure of the taxi industry in New York City.

C0401: **A multivariate randomized response model for sensitive binary data**
Presenter: **Yasuhiro Omori**, University of Tokyo, Japan
Co-authors: Amanda Chu, Mike So, Hing-yu So

A new statistical method is proposed that combines the randomized response technique, probit modeling, and Bayesian analysis to analyze large-scale online surveys of multiple binary randomized responses. We illustrate the proposed method by analyzing sensitive dichotomous randomized responses on different types of drug administration error from nurses in a hospital cluster. A statistical challenge is that nurses true sensitive responses are unobservable because of a randomization scheme that protects their data privacy to answer the sensitive questions. Four main contributions of the paper are highlighted. The first is the construction of a generic statistical approach in modeling multivariate sensitive binary data collected from the randomized response technique. The second is studying the dependence of multivariate sensitive responses via statistical measures. The third is the calculation of an overall attitude score using sensitive responses. The last one is an illustration of the proposed statistical method for analyzing administration policies that potentially involve sensitive topics which are important to study but are not easily investigated via empirical studies. A particular healthcare example of drug administration policies also presents a scientific way to elicit managerial strategies while protecting data privacy through analytics.

C0571: **Accelerated continuous space topic model for textual data**
Presenter: **Manabu Asai**, Soka University, Japan
Co-authors: Seiichi Inoue

For natural language processing, the discrete infinite logistic normal (DILN) distribution has been developed. The advantages of using Gaussian processes (GP), as in the DILN model, have been discussed. They claim that latent Gaussian noises used for stochastic kernel function in the DILN can be interpreted as a product of coordinates of words in a continuous space. They also pointed out that DILN model can be extended by accommodating additional process such as the Pitman-Yor process. For this generality, the variants of the DILN model, as the 'continuous space topic model' (CSTM), have been considered. The purpose is to improve the CSTM family using the information of semantics and style of words in the Japanese language. Our empirical result shows that the new model outperforms the existing models regarding the perplexity measure.

C0984: **High-frequency realized stochastic volatility model**
Presenter: **Toshiaki Watanabe**, Hitotsubashi University, Japan
Co-authors: Jouchi Nakajima

A new high-frequency realized stochastic volatility model is proposed. Apart from the standard daily-frequency stochastic volatility model, the high-frequency stochastic volatility model is fit to intraday returns by extensively incorporating intraday volatility patterns. The daily realized volatility calculated using intraday returns is incorporated into the high-frequency stochastic volatility model by taking account of the bias in the daily realized volatility caused by microstructure noise. The volatility of intraday returns is assumed to consist of the autoregressive process, the seasonal component of the intraday volatility pattern, and the announcement component responding to macroeconomic announcements. A Bayesian method via Markov chain Monte Carlo is developed for the analysis of the proposed model. The empirical analysis using the 5-minute returns

of Nikkei 225 index provides evidence that our high-frequency realized stochastic volatility model improves in-sample model fit and volatility forecasting over the existing models.

CO454 Room R02 MODELLING COMPLEX TIME SERIES IN ECONOMICS AND FINANCE Chair: Yang Zu

C0348: **Bootstrap based inference and probability forecasting in multiplicative error models**
Presenter: **Indeewara Perera**, University of Sheffield, United Kingdom
Co-authors: Mervyn Silvapulle

As evidenced by an extensive empirical literature, multiplicative error models (MEM) show good performance in capturing the stylized facts of nonnegative time series; examples include, trading volume, financial durations, and volatility. A bootstrap-based method is developed for producing multi-step-ahead probability forecasts for a nonnegative valued time-series obeying a parametric MEM. To test the adequacy of the underlying parametric model, a class of bootstrap specification tests is also developed. Rigorous proofs are provided for establishing the validity of the proposed bootstrap methods. The paper also establishes the validity of a bootstrap based method for producing probability forecasts in a class of semiparametric MEMs. Monte Carlo simulations suggest that our methods perform well in finite samples. A real data example involving realized volatility of the S&P 500 index illustrates the methods.

C0357: **Spatial heterogenous autoregression with varying-coefficient covariate effects**
Presenter: **Maria Kyriacou**, University of Southampton, United Kingdom
Co-authors: Zudi Lu, Peter CB Phillips, Xiaohang Ren

The traditional SARX models offer a simple way of capturing spatial interactions. Still, they have been subject to criticism owing to their several limitations, including their inability to capture spatial non-linearities and unobserved heterogeneity. We propose a spatial heterogeneous autoregressive exogenous (SHARX) model which captures for non-linearities and unobserved heterogeneity via allowing for varying-coefficients in the coefficients of the exogenous regressors (X) and the error term. The coefficients of the exogenous regressors are allowed to vary with location (s) smoothly and therefore allows to introduce spatial trends in y or heterogeneous non-linearity between X and s. Under a set of assumptions, the unknown parameters are then estimated by a profile maximum likelihood-based on a two-step procedure. First, the unknown parameters are estimated at s by local maximum likelihood estimation for a given lambda. Then the spatial profile likelihood can be defined from step 1, and the estimator of the spatial parameter is then defined as the maximum profile likelihood estimator. We assess the performance of our estimators alongside the conventional ML and GMM methods via a simulation study and an empirical application using energy data from China.

C0422: **Analyzing the social network with misspecification via double regularized GMM**
Presenter: **Chen Huang**, Aarhus University, Denmark
Co-authors: Victor Chernozhukov, Weining Wang

Social network analysis has gained significant attention recently. The identification, estimation and inference issues are intrinsically important in understanding the underlying network structure. We try to uncover the network effect with a predetermined adjacency matrix, and in addition, we allow a flexible network specification by incorporating an unobserved network structure. In particular, the unobserved network structure can be regarded as latent or misclassified network linkages. To achieve high-quality estimator for parameters in both components, we propose to estimate via a double regularized high dimensional GMM framework. Moreover, this framework also facilitates us to conduct the inference. The theory of consistency and asymptotic normality is provided with accounting for the general spatial and temporal dependency of the underlying data generating processes. Simulations demonstrate the good performance of our proposed estimation and inference procedure.

C0681: **Comparing predictive accuracy under unconditional heteroskedasticity**
Presenter: **Yang Zu**, University of Nottingham, United Kingdom
Co-authors: Steve Leybourne, Dave Harvey

The impact of unconditional heteroskedasticity on Diebold Mariano (DM) test for equal forecast accuracy is considered. We analyse the power of the DM test and propose two new powerful DM type tests by exploiting the heteroskedasticity structure in data. Empirical applications to evaluating exchange rate forecasts and the forecasts made by professionals are considered.

C1132: **Optimal probabilistic forecasts: When they work**
Presenter: **Worapree Ole Maneesoonthorn**, University of Melbourne, Australia
Co-authors: Gael Martin, David Frazier, Ruben Laoiza Maya, Andres Ramirez Hassan

Proper scoring rules are used to assess the out-of-sample accuracy of probabilistic forecasts, with different scoring rules rewarding distinct aspects of forecast performance. We re-investigate the practice of using proper scoring rules to produce probabilistic forecasts that are 'optimal' according to a given score, and assess when their out-of-sample accuracy is superior to alternative forecasts, according to that score. Particular attention is paid to relative predictive performance under misspecification of the predictive model. Using numerical illustrations, we document several novel findings within this paradigm that highlights the important interplay between the true data generating process. The assumed predictive model and the scoring rule. Notably, we show that only when a predictive model is sufficiently compatible with the true process to allow a particular score criterion to reward what it is designed to reward, will this approach to forecasting reap benefits. Subject to this compatibility; however, the superiority of the optimal forecast will be greater, the greater is the degree of misspecification. We explore these issues under a range of different scenarios and using both artificially simulated and empirical data.

CO107 Room R07 REGIME CHANGE II: FINANCE, MACRO, POLICY REGIMES Chair: Willi Semmler

C0602: **Low interest rate regime and attractive pension payouts under the benchmark approach**
Presenter: **Eckhard Platen**, University of Technology Sydney, Australia

A potential mathematically founded explanation for the currently observed low-interest rates in developed economies is given. Such low-interest rates make the production of typical pension payouts difficult when using the classical risk-neutral pricing and hedging approach. By using the more general benchmark approach, it is demonstrated how attractive pension payouts can be produced via respective dynamic asset allocation strategies.

C0878: **Credit spread, financial stress, and delayed monetary policy effectiveness**
Presenter: **Willi Semmler**, New School for Social Research, United States
Co-authors: Helmut Maurer, Pu Chen

Given the long period of expansionary monetary policies following the great recession 2008-9, many observers claim that those policies exerted their effects on the real economy through the asset market: Through the decline of financial stress, the repricing of credit risk, and declining credit spreads. To study this channel we propose a regime-switching macro model with financial stress, credit flows, and credit spreads. We study the stabilizing - destabilizing effects of the dynamics of the financial conditions on inflation and output gap. Given different regimes of financial conditions, we explore the effectiveness of conventional and unconventional monetary policies under simultaneous and delayed policy impacts. We use calibrated parameters, based on data for the Euro area, and solve the implied nonlinear dynamic system through AMPL for a finite horizon model. We find that with longer delays policies might not be able to effectively stabilize inflation and the output gap in particular if a regime switch has occurred. Though in our context the agents are forward-looking over a finite horizon, there are effects from the past that come into play with

a delay affecting real and financial variables. The possibility of asymmetric adjustments in different regimes to some long-run steady state is then empirically validated through a Multi-Regime-Cointegration-VAR (MRCIVAR) for European countries as well for the US.

C0808: **Financial stress, regime-switching and macrodynamics**
Presenter: **Pu Chen**, Melbourne Institute of Technology, Australia

Monetary responses to financial stress have recently become an important issue in macroeconomic and policy discussions in the US as well as in the EU. Two regimes of monetary responses are studied. While the fundamentals of an economy are assumed to have a long-run equilibrium, the adjustment process towards the equilibrium can be different in different regimes. During a period of deteriorated economic conditions, rates cuts are the most often applied policy responses. Therefore, rate cuts can be used as a natural regime identifier. We observe that the financial stress shocks have a large and persistent negative impact on the real side of the economy, and their impact is stronger in the non-rate-cut regime than in the rate-cut regime. A macro-foundation of such a Finance-Macro model type has been previously given. The agents can, in a finite horizon context, borrow and accumulate assets where, however, the above two scenarios may occur. The model is solved through nonlinear model predictive control (NMPC). Empirically we use a multi-regime cointegrated VAR (MRCIVAR) to study the impact of financial stress shocks and monetary policy on the macroeconomy in different countries.

C0587: **Nonlinear credit dynamics and regime switches in the output gap**
Presenter: **Francesco Simone Lucidi**, Sapienza University, Italy
Co-authors: Willi Semmler

Over the last two decades the intensity of credit standards' tightening during economic contractions has exceeded their easing during expansions among euro area banks. This mechanism is fed by the boom-bust cycle of credit that is linked to financial instability with large effects on the real economy. We build a small scale nonlinear quadratic (NLQ) model to study how credit feedback can affect the overall adjustment path of the economy towards some steady state, when the central bank solves a finite-horizon decision problem. This source of instability in the transmission mechanism is identified in a non linear VAR with time-varying volatility estimated on the euro area.

C1053: **Energy transition, asset price fluctuations,and dynamic portfolio decisions**
Presenter: **Ibrahim Tahri**, PIK (Potsdam Institute for Climate Impact Research), Germany
Co-authors: Willi Semmler, Kai Lessmann

The implications of short-termism on portfolio decisions of investors, and its potential consequences on green investments, are analyzed. We study a dynamic portfolio choice problem that contains two assets, one asset with fluctuating returns and another asset with a constant risk-free return. Fluctuating returns can arise from fossil or clean energy-related assets. Short-termism is seen to be driven by discount rates (exponential and hyperbolic) and the decision horizon of investors. We also explore the impact of the fluctuating assets returns on the fate of the portfolio, for both a deterministic and stochastic model variant, and in cases where innovation efforts are spent for fossil fuel or clean energy sources. Detailing dynamic portfolio decision in such a way may allow us for better pathways to empirical tests.

CC802 Room R03 CONTRIBUTIONS IN FINANCIAL ECONOMETRICS I Chair: Catherine Forbes

C0191: **Evaluating the underlying components of high-frequency financial data: Finite sample performance and noise**
Presenter: **Marwan Izzeldin**, Lancaster University Management School, United Kingdom
Co-authors: Rodrigo Hizmeri

The aim is to examine the finite sample properties of novel theoretical tests that evaluate the presence of: a) Brownian motion, b) jumps; c) finite vs. infinite activity jumps. In allowing for Gaussian, t-distributed, and Gaussian-T mixture noise, our Monte Carlo experiment guides a search for optimal performance across sampling frequencies. Using 100 stocks and SPY, we find that: i) a Brownian and a jump component characterize 1-min stock data; ii) Jumps should allow for both finite and infinite activity; iii) Rejection rates are time-varying, such that more jump days are usually associated with an increase of infinite jumps vis–vis finite jumps.

C0890: **Multi-period bond portfolio optimization by linear rebalancing strategy utilizing a stochastic interest rate model**
Presenter: **Yoshiyuki Shimai**, University of Tsukuba, Japan
Co-authors: Naoki Makimoto

Regardless of the asset class, it is difficult to apply multi-period dynamic portfolio optimization to real investment activity due to theoretical and structural complexity. In particular, when it comes to a bond portfolio based on a stochastic interest rate model, no empirical studies are analyzing the actual bond market. However, there exists some literature which focuses on theoretical aspects of multi-period bond portfolio optimization, such as deriving analytical solutions for optimal portfolios. Besides, a methodology to take into account realistic investment constraints has not yet been developed. We propose a new framework for multi-period bond portfolio optimization. Because bond return can be expressed as a linear combination of factors which constitute a stochastic interest rate model, we apply a linear rebalancing strategy that takes into account transaction costs in addition to self-financing constraints and short-selling constraints. Then, as empirical analysis, we conduct an investment back-test analyzing discount bonds estimated from Japanese interest-bearing government bonds. As a result, we found that the multi-period portfolio optimization represents relatively high performance compared to the single-period optimization and that the performance improves as the investment horizon and investment utilization period are extended up to a certain point.

C0730: **Bubbles on altcoins: Rush versus manipulation**
Presenter: **Rostislav Haliplii**, Universite Paris 1 Sorbonne, France

The aim is to explore the bubble effects on different crypto-currencies. Bubbles are generated by investors' urge to step-in a promising market and by price pumping trades. The main goal of the paper is to assess the presence of bubble effects in this market with customized tests able to detect the timing of various bubbles. We analyse the evolution of a representative sample crypto-currencies over time, encompassing both high and low liquidity coins. The results show that several crypto-currencies prices had episodes of rapid inflation in 2017 related to the Bitcoin bubble and a few emerging coins saw their prices pumped by speculative actions. The occurrence and the timing of bubbles in the top 50 cryptocurrencies are explored. The Sup-Augmented Dickey-Fuller and the Generalized Sup-Augmented Dickey-Fuller tests were applied for each to the full history of exchanges rates relative to the US dollars. The obtained results support our initial intuition underlining two main reasons for bubbles: the investor rush in the initial day of the coin culminating with the 2017 Bitcoin bubble and the various momentum linked to idiosyncratic factors for various coins.

C0350: **A global model of international yield curves: Regime-switching dynamic Nelson-Siegel modeling approach**
Presenter: **Takeshi Kobayashi**, NUCB Business School, Japan

The importance of examining yields at a multi-country level has been highlighted by the most financial crisis, which has shown that financial markets are globally interconnected. We have extended previous work which modeled a potentially large set of country yield curves to a regime-switching setting, proposing a global factor model in which country yield-level, slope and curvature factors may depend on global-level, slope and curvature factors as well as local factors using a monthly dataset of government bond yields for the US, Germany, Japan and the UK. The results indicate strongly that global yield-level, slope, and curvature factors do indeed exist and are economically important, accounting for a significant fraction of variation in country bond yields with interesting differences across countries. We discuss the interpretation of regime probabilities and economic variables and how the yield curve moves between two regimes. The results suggest that regimes are related to business cycles. This

results would help bonds portfolio manager to consider its country allocation decision and risk management. They contribute to the literature by assessing the degrees of the interaction of each country yield curve to a global factor in different regimes.

C0896: **Does the yield curve signal recessions? New evidence from an international panel data analysis**
Presenter: **Jean-Baptiste Hasse**, Aix-Marseille University, France

The aim is to reexamine the predictive power of the yield spread across countries and over time. Using a dynamic panel/dichotomous model framework and a unique dataset covering 13 OECD countries over a period of 45 years, we empirically show that the yield spread signals recessions. This result is robust to different econometric specifications, controlling for recession risk factors and time sampling. Using a new cluster analysis methodology, we present empirical evidence of a partial homogeneity of the predictive power of the yield spread. The results provide a valuable framework for monitoring economic cycles.

CG066 Room R06 CONTRIBUTIONS IN HIGH FREQUENCY DATA IN ECONOMICS AND FINANCE Chair: Deniz Erdemlioglu

C0440: **From previous-tick to pre-averaging: Spectra of equidistant transformations for unevenly spaced high-frequency data**
Presenter: **Vitali Alexeev**, University of Technology Sydney, Australia
Co-authors: Katja Ignatieva, Jun Chen

To convert tick-by-tick data into equidistant series, the Exponential Moving Average (EMA) sampling scheme is proposed. EMA is a parametric generalisation of the two popular methods: previous tick and pre-averaging. In essence, the proposed scheme is a spectrum of schemes that span between these two extremes. By varying a degree of the smoothing parameter, the scheme is capable of focusing on only the latest observations or favour more pronounced averaging to reduce microstructure noise. When computing the realised variance (RV) and assessing its convergence to the integrated variance (IV), existing methods have their drawbacks. Simulation study and empirical analysis demonstrate that at ultra-high sampling frequencies (10s, 20s and 30s), the EMA scheme collapses to the pre-averaging. In contrast, for lower frequencies (30-min or lower), one can rely on the previous tick sampling scheme. For frequencies ranging from 1-min to 10-min, the EMA sampling scheme must be employed to achieve reliable RV estimates.

C0976: **Periodic features and seasonality in high-frequency data**
Presenter: **Tommaso Proietti**, University of Roma Tor Vergata, Italy
Co-authors: Diego Pedregal

The advances in information technology and survey methods have increased the availability of intra-daily, daily, and weekly time series. The availability of time series observed at a high frequency, such as weekly or daily, poses new challenges that have not been properly handled in the literature. High-frequency data are relevant for producing more timely and more temporally disaggregate estimates of economic signals. However, they suffer from noise contamination, and new seasonal components are introduced. Distilling the relevant economic signals in such an environment requires the ability to handle seasonality and outliers. Robust filtering methods for preprocessing data may be required. The presentation aims at reviewing the solutions that have been provided by the literature and at exposing some of the challenges open to further research. In particular, it focuses on parametric and semiparametric models of seasonality within an unobserved components framework, where the seasonal component is estimated along with other components.

C0747: **Augmenting the realized-GARCH: The role of signed-jumps, attenuation-biases and long-memory effects**
Presenter: **Ioannis Papantonis**, Athens University of Economics and Business, Greece
Co-authors: Elias Tzavalis, Leonidas Rompolis, Orestis Agapitos

The focus is on the Realized-GARCH framework. We extend the Realized-GARCH to incorporate additional intra-day realized measures capturing different volatility asymmetry sources. We also consider a volatility-of-volatility effect, which can correct for attenuation-biases in measuring Realized Variance (RV), as well as heterogeneous terms of RV, which approximate long-memory properties of variance. These extensions are well-justified by the ongoing literature on Heterogeneous Auto-Regressive (HAR) models. Moreover, we examine the impact of allowing for skewed/leptokurtic distributions on the flexibility of the models to fit returns and variance jointly. Two main conclusions can be drawn from the results of our empirical analysis. First, the model-extensions that we suggest improve both the in- and out-of-sample performance of the model to predict RV, compared to the standard Realized-GARCH. This finding is justified by several goodness-of-fit and prediction-accuracy metrics that we report, as well as by a series of out-of-sample equal-prediction performance tests. Second, allowing for asymmetric/fat-tailed return distributions also seems to play a crucial role in the accurate filtering of the innovations, and consistently helps the identification of the parameters of the volatility process. This further enhances the prediction performance of all the augmented GARCH-type specifications that we consider in our analysis.

C0426: **A state space approach based on pre-averaging sampling scheme for estimation of integrated variance**
Presenter: **Jun Chen**, UNSW Sydney , Australia
Co-authors: Vitali Alexeev, Katja Ignatieva

A new state-space model is proposed to estimate the Integrated Variance (IV) in presence of microstructure noise. Applying the pre-averaging sampling scheme to the irregularly spaced high frequency data, we derive equidistant efficient price approximations to calculate the noise-contaminated RV (NCRV), which is used as the IV estimator. The theoretical properties of the new volatility estimator are illustrated and compared with those of the realized volatility. We highlight the robustness of the new estimator to market microstructure noise. The empirical illustration features EUR/USD exchange rate and provides evidence of an efficient performance in volatility forecasting at very high sampling frequency.

C0627: **Realized principal component analysis: A pre-averaging approach**
Presenter: **Francesco Benvenuti**, Aarhus University, Denmark
Co-authors: Kim Christensen, Bezirgen Veliyev

A Realized Principal Component Analysis (RPCA) theory robust to noise is proposed, using a pre-averaging approach. The RPCA is the high-frequency extension of the classic PCA: in this setting noise always contaminates price observations. Hence, we first obtain a consistent estimator of the spot covariance matrix employing the well-known pre-averaging technique. Then, we derive the realized eigenvalue, eigenvectors and principal component estimators for this case, building on a recent theory about volatility functional estimation. We conclude by providing simulation results in different noise scenarios, and we discuss how those estimators work in practice.

EO706 Room R08 FRONTIERS IN POPULATION GENETICS Chair: Ritabrata Dutta

E0250: **Estimating recombination from population genetic data**
Presenter: **Andreas Futschik**, JKU Linz, Austria

Recombination leads to the reshuffling of genetic material during meiosis. The population recombination rate is an important parameter which informs about the effect of recombination on populations. It is defined as four times the recombination rate per base pair and generation multiplied with the effective population size. Estimating the population recombination rate, as well as the identification of hotspots, is of considerable interest. Therefore several methods have been proposed for this purpose that are often computationally quite expensive. We discuss how existing estimates can be improved by optimizing the bias-variance trade-off. We also propose a new fast and accurate estimate that uses relevant summary statistics in a regression model.

E0284: **Bayesian inference of PolII pausing dynamics over exclusion processes**
Presenter: **Massimo Cavallaro**, University of Warwick, United Kingdom
Co-authors: Yuexuan Wang, Daniel Hebenstreit, Ritabrata Dutta

Transcription is a complex phenomenon that allows the conversion of genetic information into phenotype through an enzyme called PolII, which erratically moves along and scans the DNA template. We perform Bayesian inference over a paradigmatic mechanistic model of non-equilibrium statistical physics, i.e., the asymmetric exclusion processes (ASEP) in mean-field approximation, assuming a Gaussian process prior for the PolII progression rate as a latent variable. Our framework allows us to infer the speed of PolIIs during transcription, given their spatial distribution whilst avoiding the explicit inversion of the system's dynamics. The results may have implications for the understanding of gene expression and biological noise.

E0540: **Inference of natural selection from allele frequency time series data using exact simulation techniques**
Presenter: **Paul Jenkins**, University of Warwick, United Kingdom
Co-authors: Jaromir Sant, Jere Koskela, Dario Spano

A standard problem in population genetics is to infer evolutionary and biological parameters such as the effective population size, mutation rates, and strength of natural selection from DNA samples extracted from a contemporary population. That all samples come only from the present-day has long been known to limit statistical inference; there is potentially more information available if one also has access to ancient DNA so that inference is based on a time-series of historical changes in allele frequencies. We introduce a Markov Chain Monte Carlo method for Bayesian inference from allele frequency time-series data based on an underlying Wright-Fisher diffusion model of evolution. The chief novelty is that we show this method to be exact in the sense that it is possible to augment the state space explored by MCMC with the unobserved diffusion trajectory, even though the transition function of this diffusion is intractable. Through careful design of a proposal distribution, we describe an efficient method in which updates to the trajectory and accept/reject decisions are calculated without error. We illustrate the method on data capturing changes in coat colour during the domestication of the horse.

EO421 Room R11 EXPERIMENTAL DESIGN AND DATA ANALYSIS Chair: Kalliopi Mylona

E0951: **Echelon designs, Hilbert series and Smolyak grids**
Presenter: **Hugo Maruri**, QMUL, United Kingdom
Co-authors: Henry Wynn

Echelon designs were first described in 2000. These designs are defined for continuous factors and include, amongst others, factorial designs. They have the appealing property that the saturated polynomial model associated with it mirrors the geometric configuration of the design. Perhaps surprisingly, the interpolators for such designs are based upon the Hilbert series of the monomial ideal associated with the polynomial model and thus the interpolators satisfy properties of inclusion-exclusion. Echelon designs are quite flexible for modelling and include the recently developed designs known as Smolyak sparse grids. We present the designs, describe their properties and show examples of application.

E0989: **Probability-based optimal designs to minimise separation**
Presenter: **Steven Gilmour**, KCL, United Kingdom
Co-authors: Mohammad Lutfor Rahman

Separation is a common problem in models with binary responses when one or more covariates predict perfectly some binary outcome. Separation is observed during the fitting of logistic models where at least one parameter estimate diverges to infinity. The separation problem leads to convergence difficulties as well as the non-existence of maximum likelihood estimates (MLEs) of model parameters. Researchers are usually advised to deal with separation either by undertaking post hoc data adjustment or by estimation corrections. However, apart from these solutions, the probability of separation arising can be greatly reduced by appropriate design of the experiment. A simple method for doing this is introduced, using newly developed Ps- and DPs- optimality criteria at the design stage. A simple result shows that we should avoid exact replicates, but beyond this, reducing separation and improving parameter estimation are in conflict. We suggest a compound criterion to compromise between these two objectives. Simulation results confirm that the designs produced to achieve the required improvements, and they can be recommended for practical use.

E1009: **Multi-fidelity statistical modelling for molecular crystal structure prediction**
Presenter: **Olga Egorova**, University of Southampton, United Kingdom
Co-authors: Roohollah Hafizi, David Woods, Graeme Day

Structural polymorphism occurs when crystallising the same molecule results in obtaining multiple (up to tens of thousands) of solid forms which vary in terms of the associated lattice energies. Obtaining reliable energy evaluations is of great importance, as energy differences are linked to the differences in physical and chemical properties of the structures which affect, for example, the suitability and safety of a pharmaceutical compound. Computational methods for crystal structure prediction (CSP), which would allow for reliable energy evaluations are highly demanding in terms of computational costs, making their direct application for all trial structures infeasible. We employ multi-fidelity Bayesian Gaussian process modelling to combine different levels of computational methods to obtain predictions for the highest level at much lower costs. We assess the uncertainties of the obtained energy predictions together with their propagations in energy rankings. The approach for energy surface optimisation is also considered.

EO169 Room R12 CLUSTERING OF COMPLEX DATA STRUCTURE Chair: Maria Brigida Ferraro

E0693: **A novel structure-based approach for multivariate time series clustering**
Presenter: **Angel Lopez Oriona**, Universidade da Coruña, Spain
Co-authors: Jose Vilar

Clustering of multivariate time series (MTS) is a central problem in data mining with many applications. Frequently, the clustering target is to identify groups of MTS generated by the same multivariate stochastic process. Most of the approaches to address this problem include a prior

step of dimensionality reduction which may result in a loss of information on the structural relationships of the MTS or consider dissimilarities based on correlations and cross-correlations, but ignoring the serial dependence structure. We propose a novel approach to measure dissimilarity between MTS aimed at jointly measuring both cross-sectional and serial dependence. Each MTS is characterized by a set of matrices of estimated quantile cross-spectral densities, where each matrix corresponds to an arbitrary pair of quantile levels. Then the dissimilarity between every couple of MTS is evaluated by comparing their estimated quantile cross-spectral densities, and the pairwise dissimilarity matrix is taken as a starting point to develop a partitioning around medoids (PAM) algorithm. Since the quantile-based cross-spectra capture dependence in quantiles of the joint distribution, the proposed metric has a high capability to discriminate between high-level dependence structures. An extensive simulation study shows that our clustering procedure outperforms a wide range of alternative methods, besides being computationally efficient. A real data application illustrates the usefulness of our approach.

E0798: **Biclustering ordinal data through a model-based approach**
Presenter: **Monia Ranalli**, Sapienza University of Rome, Italy
Co-authors: Francesca Martella

A finite mixture model to simultaneously cluster the rows and columns of a two-mode ordinal data matrix is proposed. Following the Underlying Response Variable (URV) approach, the observed variables are considered as a discretization of latent continuous variables distributed as a mixture of Gaussians. To introduce a partition of the P variables within the g-th component of the mixture, we adopt a factorial representation of the data where a binary row stochastic matrix, representing variable membership, is used to cluster variables. In this way, we associate a component in the finite mixture to a cluster of variables and define a bicluster of units and variables. The number of clusters of variables (and therefore the partition of variables) may vary with clusters of units. Due to the numerical intractability of the likelihood function, estimation of model parameters is based on composite likelihood (CL) methods. It essentially reduces to a computationally efficient Expectation-Maximization type algorithm. The performance of the proposed approach is discussed in both simulated and real datasets.

E0935: **Fuzzy clustering of networks**
Presenter: **Ilaria Bombelli**, Sapienza University of Rome, Italy
Co-authors: Ichcha Manipur, Mario Guarracino, Maria Brigida Ferraro

Networks represent a powerful model to describe problems and applications in various fields, such as economics, science and technology. The focus is on the fuzzy clustering of networks. In detail, we provide computational procedures to look for clusters of networks, where each network represents an object. Our proposal is based on the Non-Euclidean Fuzzy Relational Clustering (NEFRC) algorithm. Since the algorithm requires as input a distance matrix, we need some specific measures of distance between networks. First of all, we represent each network using probability distributions (Node Distance Distribution and Transition Matrices) to obtain a matrix representation. Then we use a measure of dissimilarity between probability distributions, and we get the networks distance matrix the NEFRC algorithm requires as input. We check the adequacy of the proposals through simulations and real-case studies.

EO498 Room R13 PUBLIC POLICY ANALYSIS AND MACHINE LEARNING I Chair: Michela Bia

E0759: **A regression discontinuity design for ordinal running variables: Evaluating central bank purchases of corporate bonds**
Presenter: **Andrea Mercatanti**, Banca d'Italia, Italy
Co-authors: Taneli Makinen, Andrea Silvestrini, Fan Li

Regression discontinuity (RD) is a widely used quasi-experimental design for causal inference. In the standard RD, the treatment assignment is determined by a continuous pretreatment variable (i.e., running variable) falling above or below a pre-fixed threshold. Recent applications increasingly feature ordered categorical or ordinal running variables, which pose challenges to RD estimation due to the lack of a meaningful measure of distance. An RD approach is proposed for ordinal running variables under the local randomization framework. The proposal first estimates an ordered probit model for the ordinal running variable. The estimated probability of being assigned to treatment is then adopted as a latent continuous running variable and used to identify a covariate-balanced subsample around the threshold. Assuming local unconfoundedness of the treatment in the subsample, an estimate of the effect of the program is obtained by employing a weighted estimator of the average treatment effect. Two weighting estimators—overlap weights and ATT weights—, as well as their augmented versions, are considered. We apply the method to evaluate the causal effects of the corporate sector purchase programme (CSPP) of the European Central Bank, which involves large-scale purchases of securities issued by corporations in the euro area. We find a statistically significant and negative effect of the CSPP on corporate bond spreads at issuance.

E0819: **Causal mediation analysis with double machine learning**
Presenter: **Martin Huber**, University of Fribourg, Switzerland
Co-authors: Martin Spindler, Henrika Langen, Lukas Laffers, Helmut Farbmacher

Causal mediation analysis is combined with double machine learning for a data-driven control of observed confounders in a high-dimensional setting. The average indirect effect of a binary treatment and the unmediated direct effect are estimated based on efficient score functions, which are robust 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. We demonstrate that the effect estimators are asymptotically normal and root-n consistent under specific regularity conditions and provide a simulation study as well as an application to the National Longitudinal Survey of Youth.

E1180: **Causal effects with hidden treatment diffusion over partially unobserved networks**
Presenter: **Costanza Tortu**, IMT School for Advanced Studies Lucca, Italy
Co-authors: Irene Crimaldi, Fabrizia Mealli, Laura Forastiere

In randomized experiments where some units are randomly assigned to a treatment, interactions between units might generate a treatment diffusion process. For instance, if the intervention of interest is an information campaign realized through a video or a flyer, some treated units might share the treatment with their friends. Such a phenomenon, which is usually hidden, causes a misallocation of individuals in the two treatment arms: some of the initially untreated units might have actually received the treatment by diffusion. This circumstance, in turn, might introduce a bias in the estimate of the causal effect of the intervention. Inspired by a recent field experiment on the effect of different types of school incentives aimed at encouraging students to attend cultural events, we present a novel approach to deal with a hidden diffusion process, in the presence of a partially unknown network structure. We address the issue of a partially unobserved network by imputing the presence (or the absence) of missing ties, using random forests. Then, we develop a simulation-based sensitivity analysis that assesses the robustness of the estimates against the possible presence of a treatment diffusion. We simulate several diffusion scenarios within a plausible range of sensitivity parameters, and we compare the treatment effect, which is estimated in each scenario with the one that is obtained while ignoring the diffusion process.

EO650 Room R14 DATA-CENTRIC ENGINEERING: A NEW CHALLENGE FOR STATISTICIANS Chair: Dirk Fromme

E1041: **Understanding complex fluids with data-centric rheology**
Presenter: **Jack Hale**, University of Luxembourg, Luxembourg

The purpose is to show some recent applications of statistical methods, such as non-parametric regression and model selection, in understanding the

rheology of complex viscoelastic and non-Newtonian fluids. A perspective on the success and difficulties of applying these methods to challenging problems in the physical sciences will be given.

E1125: **Manifold MCMC methods for inference in inverse problems with highly informative observations**
Presenter: **Matt Graham**, Newcastle University, United Kingdom
Co-authors: Alexandre Thiery, Khai Xiang Au

Inverse problems - inferring the configuration of a model of a physical system given observations - abound in engineering settings. Typically, the inverse problem is ill-posed. In such settings, Bayesian methodology offers a principled approach for combining prior knowledge with observations to infer the posterior distribution of plausible configurations of the model. A particularly challenging setting is where the data are highly informative with a large signal-to-noise ratio. Such observations lead to a posterior which concentrates around a lower-dimensional manifold embedded in the model configuration space. This high-fidelity observation regime is common in engineering settings where often the measurement process is carefully designed to minimise the effects of noise. Existing Markov chain Monte Carlo (MCMC) methods struggle in this regime, requiring an increasing computational effort as the signal-to-noise ratio grows. We will present a strategy that transforms the original sampling problem into the task of exploring a distribution supported on a manifold embedded in a higher-dimensional space. In contrast to the original posterior, this lifted distribution remains diffuse in the limit of vanishing observation noise. By leveraging the geometry of this lifted posterior, we propose an MCMC method which remains efficient as the signal-to-noise ratio increases, allowing complex simulator models to be efficiently calibrated against high-fidelity observations.

E1182: **Bayesian polynomial chaos in multi-fidelity modelling**
Presenter: **Pranay Seshadri**, Imperial College London, United Kingdom
Co-authors: Andrew Duncan

Bayesian polynomial chaos, a Gaussian process analogue to polynomial chaos, will be introduced. Polynomial chaos represents a set of methodologies for delivering efficient aleatory uncertainty estimates for computer models. It has garnered significant industrial uptake within engineering, having seen applications in aerospace, mechanical, civil, geothermal and wind sectors. We argue why our Bayesian re-formulation of polynomial chaos is necessary and proceed to define it mathematically. The thrust will be on how Bayesian polynomial chaos is tailored for multi-fidelity uncertainty quantification, where one has to negotiate data from multiple models of varying fidelity and associated experimental data. An application of the proposed methodology on a gas turbine is presented.

EO668 Room R15 MODELING UNCERTAINTY AND VAGUENESS IN DECISION MAKING AND ECONOMICS Chair: Davide Petturiti

E0731: **Relation between volatility and returns through a quantile fuzzy regression model**
Presenter: **Maria Letizia Guerra**, , Italy
Co-authors: Luciano Stefanini

The purpose is to analyze the nature of the relation of the pair volatility and return in the particular case of CBOE VIX and S&P 500 index. In particular, the S&P500 returns time series is modelled through fuzzy-valued functions, whose level-cuts are interpreted in the framework of expectile and quantile fuzzy regressions which are built by defining fuzzy-valued expectile (L2-norm) and quantile (L1-norm) extensions of the F-transforms. Since in the whole time period the relationship between VIX and S&P500 returns changes dramatically, we introduce the clustering of the data into subsets to significantly improve the quality of fitting; the clustering is applied only when a preliminary evaluation test based on Kendall and Spearman correlation verifies its e efficacy. A forecasting methodology is proposed based on specific forms of local trends such as parametric exponential functions which we prove to be more suitable and stable for extrapolation than polynomials. We show how it is possible to forecast S&P500 returns to time $T+M$ (M-steps ahead) by having available volatility and returns observations till time T.

E0779: **Processing distortion models: A comparative study**
Presenter: **Enrique Miranda**, University of Oviedo, Spain
Co-authors: Ignacio Montes, Sebastien Destercke

When dealing with sets of probabilities, distortion or neighbourhood models are convenient, practical tools, as very few parameters determine them: an initial probability distribution pr_0, a distortion factor $\delta > 0$ and a specific distortion procedure. The different choices have led to several different families of neighbourhood models, with applications in robust statistics or machine learning. We compare the performance of several distortion models under several processing procedures. First of all, we study their behaviour when merging different distortion models quantifying uncertainty on the same quantity using conjunction, disjunction or convex mixtures. Secondly, we investigate whether the marginal credal sets of a distortion model are also members of the same family, as well as the procedure for determining a global model from marginal ones using independence or natural extension. The analysis is made for six different families of distortion models: the pari-mutuel, epsilon-contamination, constant odds-ratio, total variation, Kolmogorov and L_1.

E1073: **Envelopes of equivalent martingale measures in an n-nomial market model**
Presenter: **Andrea Cinfrignini**, University of Rome - La Sapienza, Italy
Co-authors: Davide Petturiti, Barbara Vantaggi

An n-nomial market model over one time period is considered, composed by a risky asset and a risk-free bond. It is well-known that such a model, though arbitrage-free, is incomplete for $n > 2$, as it gives rise to a family of equivalent martingale measures. In general, given a contingent claim on the risky asset, the approach under incompleteness is to choose one of the equivalent martingale measures in the class in order to arrive at a unique no-arbitrage price for the contract. A different approach is to work with the entire class or with a suitable subclass: in this case, we get an interval of prices. Here, we provide a characterization of the lower envelope of the class of equivalent martingale measures, casting it in the Dempster-Shafer theory of evidence. We further introduce a generalized no-arbitrage principle and investigate how to obtain a pricing functional from the lower envelope which is generalized-arbitrage-free.

EO081 Room R21 ADVANCES IN SURVIVAL AND RELIABILITY II Chair: Mariangela Zenga

E0283: **Bayesian inference for systems and network reliability using the survival signature**
Presenter: **Simon Wilson**, Trinity College Dublin, Ireland
Co-authors: Louis Aslett, Frank Coolen

The concept of survival signature is an alternative to the signature for reliability quantification of a system that is suitable for systems with multiple types of component. This allows the use of the survival signature for estimation of the reliability of networks. We present the use of the survival signature for reliability quantification of systems and networks from a Bayesian perspective, using data on tested components that are exchangeable with those in the actual system or network of interest. These data consist of failure times and possibly right censoring times. A nonparametric as well as a parametric approach are described.

E0400: **Design of a periodic inspection policy in heterogeneous systems with two types of components**
Presenter: **Maria Lucia Bautista Barcena**, Universidad de Extremadura, Spain
Co-authors: Inmaculada Torres Castro

A heterogeneous system consisting of monitored and non monitored components is analyzed. Monitored components are subject to a degradation

process (following a gamma process) and they fail when their degradation level exceeds a corrective threshold. Condition-based maintenance is applied to reduce the impact of the failures on them. Non monitored components are subject to sudden failures. They can only be maintained correctively upon failure. Maintenance team takes a while to repair the system. The system is inspected periodically every T time units. In these inspection times, if the degradation level of a monitored component reaches a certain preventive threshold, this component is replaced by a completely new one. Also, an opportunistic maintenance policy is implemented to take advantage of system failures and perform preventive maintenance on some components if necessary. Each maintenance task implies a certain cost and each monitored component is assumed to provide a reward which decreases when the deterioration level of the component increases. Assuming infinite time span, the expected cost rate of the system is minimised through the optimization of the preventive thresholds and time between inspections. Meta-heuristic algorithms such as genetic algorithms or pattern search, combined with typical Monte Carlo simulation methods, are used to compute the maintenance cost of the system.

E0671: **Phase-type distributions in a vector Markov process to analyse random telegraph noise in resistive memories**
Presenter: **Christian Acal**, University of Granada, Spain
Co-authors: Juan Eloy Ruiz-Castro, Ana Maria Aguilera, Juan Bautista Roldan

Advanced statistical techniques are key tools to model complex physical and engineering problems in many different areas of expertise, such as the field of Resistive Random Access Memories (RRAMs). One of the most important aspects to consider, prior to the massive industrialization, is the Random Telegraph Noise (RTN). This issue is a great concern because it can affect the correct operation of a device. A device can emit signals originated thanks to disturbances produced by several traps that provoke current fluctuations. This process can be represented as a process that evolves over time, going through multiple states. In this line, a vector Markov process, by considering macro-states in order to analyse, model and study the evolution, is introduced. Multiple measures of interest are worked out. So far, the usual statistical analysis performed on the sojourn times makes use of the exponential distribution; however, sometimes its fit is not accurate and therefore, another via must be considered. In this point, we propose a novel approach based on Phase-Type Distributions (PHD) where each stage of the device is a macro-state, in a way that the sojourn time distribution for each macro-state is PH distributed.

EO115 Room R22 STATISTICS AND DECISION MAKING FOR THE COVID-19 PANDEMIC Chair: Raimund Kovacevic

E0753: **A distributed optimal control epidemiological model: Parameter identification and optimization**
Presenter: **Raimund Kovacevic**, Vienna University of Technology, Austria
Co-authors: Vladimir Veliov, Nikolaos Stlianakis

A distributed optimal control epidemiological model is presented that describes the dynamics of an epidemic with social distancing as a control policy. The model belongs to the class of continuous-time ordinary/partial differential equation models but has an important novel feature. The core model - a single integral equation - does not explicitly involve transition rates between compartments. Instead, it is based on statistical information on the disease status of infected individuals, depending on the time since infection. The model is especially relevant for the coronavirus 2019 (COVID-19) disease in which infected individuals are infectious before onset of symptoms during a relatively long incubation period.

E0845: **Understanding the relationship between the uptake of Covid-19 testing and deprivation**
Presenter: **Doris Behrens**, Aneurin Bevan University Health Board / Cardiff University, United Kingdom
Co-authors: Dan Davies, William Beer, Daniel Westwood, Eryl Powell

Aneurin Bevan University Health Board (ABUHB) provides Primary, Secondary, Community and Mental Health Services for 630,000 people in South-East Wales. ABUHB hosts both the nation's most affluent and most deprived areas and the health board is part of the region that was (and still is) one of the UK's Covid-19 hot spots. Understanding the relationship between local Covid-19 testing rates and deprivation is thus of utmost importance for decision-making and the planning of localised Public Health interventions to confine the spread of the disease in this heterogeneous society. The findings derived by a combination of Geographical Analytics, Statistical Process Control and Inferential Statistics busted some myths concerning the uptake of testing. Moreover, they provided intelligence about where to set up targeted interventions to overcome inequalities and inadequacies in access to services. In response, ABUHB established sampling facilities which allow low vehicle ownership groups to walk up to the testing site. Areas with large populations of foreign descent received bespoke support in a range of languages. Additionally, Mobile Testing Units (MTU) have been deployed throughout ABUHB to improve further the accessibility of testing within areas of high prevalence. As a next step, outreach teams will support those fellow-citizens which are traditionally hardest to reach.

E0852: **An indirect method to monitor the fraction of people ever infected with COVID-19: An application to the United States**
Presenter: **Miguel Sanchez-Romero**, Wittgensteincentre for Demography and Human Capital, Austria
Co-authors: Alexia Prskawetz, Vanessa di Lego, Bernardo Lanza Queiroz

The number of COVID19 infections is key for accurately monitoring the pandemics. However, due to differential testing policies, asymptomatic individuals and limited large-scale testing availability, it is challenging to detect all cases. Seroprevalence studies aim to address this gap by retrospectively assessing the number of infections, but they can be expensive and time-intensive. We propose a complementary approach that combines estimated (1) infection fatality rates (IFR) using a Bayesian melding SEIR model with (2) reported case-fatality rates (CFR) to estimate the fraction of people ever infected and detected indirectly. We apply the technique to the U.S due to their remarkable regional diversity and because they count with almost a quarter of all global confirmed cases and deaths. We obtain that the IFR varies from 1.25% (0.39-2.16%, 90% CI) in Florida, the most aged population, to 0.69% in Utah (0.21-1.30%, 90% CI), the youngest population. By September 8, 2020, we estimate that at least five states have already a fraction of people ever infected between 10 and 20 %. The state with the highest estimated fraction of people ever infected is New Jersey, with 17.3% (10.0, 55.8, 90% CI). The results also indicate that with a probability of 90% the fraction of detected people among the ever infected since the beginning of the epidemic has been less than 50% in 15 out of the 20 states analyzed.

EC799 Room R20 CONTRIBUTIONS IN APPLIED STATISTICS II Chair: Sonja Greven

E0899: **A new R-package for the analysis of the fisheries population under uncertainty**
Presenter: **Marta Cousido Rocha**, University of Vigo, Spain
Co-authors: Cousido-Rocha Marta, Santiago Cervino, Maria Grazia Pennino

The analysis of the dynamic of a population has become a fundamental tool in ecology, conservation, biology, and particularly, in fisheries science to assess the status of exploited resources. Uncertainty is an inherent component in fishery systems that makes difficult taking management decisions. Hence, we have developed our new R package Rfishpop (available on https://github.com/IMPRESSPROJECT/Rfishpop) to deal with uncertainty for analyzing exploited populations in R. This package addresses such aims by implementing a completed Management Strategy Evaluation (MSE) cycle. MSE is a tool that scientists can use to simulate the behaviour of a fisheries system and allow them to test whether potential management procedures can achieve pre-agreed management objectives. We describe all the interlinked model structures in MSE and its implementation. Furthermore, we provide the main conclusions and a discussion about open issues.

E0785: **Depression and stigmatization: A multilevel analysis for EU countries**
Presenter: **Patricia Moreno**, Universidad de Cantabria, Spain
Co-authors: Ana Fernandez, Juan Manuel Rodriguez-Poo

The purpose is to investigate the cross-country differences in the effect that depression has on working hours. The main objective is to provide a comparable framework for a population with depression though multilevel analysis, taking into account the existence of sample selection effects and

an endogeneity problem due to the presence of the variable depression. Depression contributes to early retirement, and some working conditions are associated with an increase in the risk of depressive symptoms. We use data from regular panel waves of Survey of Health, Ageing and Retirement in Europe (SHARE). This panel database is constituted by microdata information on health, socio-economic status and social factors of more than 120000 observations. The survey covers 27 European countries and Israel. We conclude that the average hours worked per individual is higher in countries with less population suffering from depression.

E0673: **Indirect techniques to study gender violence**
Presenter: **Beatriz Cobo**, University of Granada, Spain
Co-authors: Maria del Mar Rueda

Social research is faced with difficulties in finding safe and truthful answers since respondents offer a respondent-based on what is socially acceptable or hide information by giving false answers or do not respond. Indirect techniques were developed to be reliable when faced with sensitive questions. Since its creation, many techniques have been developed, specifically, we are going to focus on the item count technique and the crosswise technique to carry out a study on gender violence. Gender-based violence is an act of violence for reasons of sex that has or may result in physical, sexual or psychological harm or suffering for women, as well as threats of such acts, coercion or arbitrary deprivation of liberty, whether they occur in public life or private life. The World Health Organization indicates that gender violence is a priority problem in public health and that it must be addressed from different areas. For this reason, the prevention of these situations is crucial, and numerous studies are carried out that reflect the current situation to take measures. We focus on physical, emotional and sexual violence, and the estimates obtained through indirect techniques are compared with those obtained from direct questioning, and regression models are studied.

EG012 Room R16 CONTRIBUTIONS IN COMPUTATIONAL AND METHODOLOGICAL STATISTICS Chair: Enea Bongiorno

E0580: **Spectral clustering on spherical coordinates under the degree corrected stochastic blockmodel**
Presenter: **Francesco Sanna Passino**, Imperial College London, United Kingdom
Co-authors: Nicholas Heard, Patrick Rubin-delanchy

Spectral clustering is a popular method for community detection in networks. The nodes are clustered on a low dimensional representation of the graph, resulting from a truncated spectral decomposition of the adjacency matrix or one of its regularised versions. Estimating the number of communities and the dimensionality of the reduced latent space is crucial for a good performance of the spectral clustering algorithm in estimating stochastic blockmodels. Additionally, real-world networks often present heterogeneous within-community degree distributions, for example, in cybersecurity applications. This property is addressed within community detection by the degree corrected stochastic blockmodel. A novel model-based method for simultaneous and automated selection of the number of communities and latent dimensionality in spectral clustering under the degree-corrected stochastic blockmodel is proposed. The method is based on a transformation to spherical coordinates of the spectral embedding, and on a novel modelling assumption in the transformed space, then embedded into a model selection framework for the number of communities and latent dimensionality. Results show improved performance over competing methods on simulated and real-world computer network data.

E0596: **Latent group structure and regularized regression**
Presenter: **Konstantinos Perrakis**, Department of Mathematical Sciences, Durham University, United Kingdom
Co-authors: Thomas Lartigue, Frank Dondelinger, Sach Mukherjee

Regression models generally assume that the conditional distribution of response Y given features X is the same for all samples. For heterogeneous data with distributional differences among latent groups, standard regression models are ill-equipped, especially in large multivariate problems where hidden heterogeneity can easily pass undetected. To allow for robust and interpretable regression modeling in this setting, we propose a class of regularized mixture models that couples together both the multivariate distribution of X and the conditional $Y|X$. This joint modeling approach offers a novel way to deal with suspected distributional shifts, which allows for automatic control of confounding by latent group structure and delivers scalable, sparse solutions. Estimation is handled via an expectation-maximization algorithm, whose convergence is established theoretically. We illustrate the key ideas via empirical examples.

E1196: **Integrate dietary assessments with biomarker measurements in aetiological models**
Presenter: **Marta Pittavino**, Geneva School of Economics and Management (GSEM), Research Center for Statistics (RCS), Switzerland
Co-authors: Pietro Ferrari, Martyn Plummer

In nutritional epidemiology, self-reported assessments of dietary exposure are prone to measurement errors, which is responsible for bias in the association between dietary factors and risk of disease. In this study, self-reported dietary assessments were complemented by biomarkers of dietary intake. Dietary and serum measurements of folate and vitamin-B6 from two nested case-control studies within the European Prospective Investigation into Cancer and Nutrition (EPIC) study were integrated into a Bayesian model to explore the measurement error structure of the data, and relate dietary exposures to risk of site-specific cancer. A Bayesian hierarchical model was developed, which included: 1) an exposure model, to define the distribution of unknown true exposure (X); 2) a measurement model, to relate observed assessments, in turn, dietary questionnaires (Q), 24-hour recalls (R) and biomarkers (M) to X measurements; 3) a disease model, to estimate exposures/cancer relationships. The marginal posterior distribution of model parameters was obtained from the joint posterior distribution, using Markov Chain Monte Carlo (MCMC) sampling techniques in JAGS. This challenging developmental work will be described, together with preliminary findings.

CO031 Room R02 TOPICS IN TIME SERIES AND PANEL DATA ECONOMETRICS Chair: Indeewara Perera

C0192: **Semi-parametric single-index predictive regression models with cointegrated regressors**
Presenter: **Weilun Zhou**, University of Cambridge, United Kingdom
Co-authors: Jiti Gao, David Harris, Hsein Kew

The estimation of a semi-parametric single-index predictive regression model in the presence of cointegration among the multivariate predictors is considered. This model is useful for predicting financial asset returns, whose behaviour is compatible with a stationary series, when the multiple predictors are nonstationary, and also allows for nonlinear predictive relationships. The single-index specification, which contains the cointegrated predictors, not only solves the problem of unbalance in the predictive regression, but also avoids the problem of the curse of dimensionality associated with fully nonparametric multivariate models. An orthogonal series expansion is used to approximate the unknown link function for the single-index component. We consider the constrained nonlinear least squares estimator of the single-index (or the cointegrating) parameters and the plug-in estimator of the link function, and derive their asymptotic properties. In an empirical application, we find some evidence of in-sample nonlinear predictability of U.S. stock returns using cointegrated predictors. We also find that the single-index model in general produces better out-of-sample forecasts than both the prevailing mean model and the linear predictive regression model.

C0198: **Issues in the estimation of misspecified models of fractionally integrated processes**
Presenter: **Kanchana Nadarajah**, University of Sheffield, United Kingdom
Co-authors: Gael Martin, Donald Poskitt

The aim is to provide a comprehensive set of new theoretical results on the impact of mis-specifying the short run dynamics in fractionally integrated processes. We show that four alternative parametric estimators - frequency domain maximum likelihood, Whittle, time domain maximum likelihood and conditional sum of squares - converge to the same pseudo-true value under common mis-specification, and that they possess a common asymptotic distribution. The results are derived assuming the true data generating mechanism is a fractional linear process driven by a martingale

difference innovation. A completely general parametric specification for the short run dynamics of the estimated (mis-specified) fractional model is considered, and with long memory, short memory and antipersistence in both the model and the data generating mechanism accommodated. An existing line of research on mis-specification in fractional models is extended. It also complements a range of existing asymptotic results on estimation in correctly specified fractional models. Open problems in the area are the subject of the final discussion.

C0456: **Time-varying panel data models with additive factor structure**
Presenter: **Fei Liu**, Nankai University, China

A nonparametric panel data model with time-varying regression coefficients and an additive factor structure is considered. This model is motivated by some explored features of real data from economics and finance. In the model, factor loadings are unknown functions of observable variables which can capture time-variant and heterogeneous covariate information. We propose a profile marginal integration (PMI) method to estimate unknown coefficient functions, factors and their loadings jointly in a single step. The asymptotic distributions for the proposed profile estimators are established. Our research fills the gap of insufficient discussions on the factors and loadings' asymptotic properties. The finite sample performance of our estimators is assessed by both simulations and empirical studies on US stock return data, which demonstrate the advantages in modelling and estimation approach in practice.

CO191 Room R03 FINANCIAL ECONOMETRICS: INTRINSIC TIME, VOLATILITY ESTIMATION, JUMP TESTING Chair: Ingmar Nolte

C0495: **Volatility estimation and sampling efficiency: An intrinsic time approach**
Presenter: **Yifan Li**, The University of Manchester, United Kingdom
Co-authors: Ingmar Nolte, Sandra Nolte

The concept of intrinsic time sampling (ITS) is developed for a semimartingale, which samples the semimartingale whenever its path triggers some homogenous stopping rule. Based on the ITS scheme, we propose the intrinsic time volatility (ITV) estimator, which is a class of consistent and jump-robust integrated variance (IV) estimators. When the path of the semimartingale is available, we show that the ITS based realized variance and the ITV estimators have smaller asymptotic variances relative to their calendar time-based counterparts under a common sampling frequency. This is driven by the fact that the ITS scheme contains more information about the integrated variance, a concept which we formalize as the sampling efficiency of sampling schemes. We derive explicit bias-correction to the ITV estimators when the semimartingale is observed discretely in time, which provides a simple and effective variance reduction technique for IV estimation using sparsely sampled data.

C0497: **Testing for jumps: An increment censoring approach in boundary-hitting intrinsic time**
Presenter: **Shifan Yu**, Lancaster University, United Kingdom
Co-authors: Yifan Li, Ingmar Nolte, Sandra Nolte

A novel nonparametric test is proposed to determine whether finite-activity jumps are present in a discretely observed price process or not. We use the concept of censored realized variation for the observations sampled at hitting times with respect to a symmetric double barrier to construct our test statistics for a univariate Ito semimartingale. The test statistics diverge to infinity if jumps are present and have a normal distribution otherwise. We use the Monte Carlo simulations to investigate the performance of our new tests in a range of empirically relevant scenarios and comparing them with other nonparametric jump tests based on calendar time sampling. We finally present the empirical results using the real-world high-frequency financial data.

C0519: **Separate noise and jumps: A price duration approach**
Presenter: **Seok Young Hong**, Lancaster University Management School, United Kingdom
Co-authors: Oliver Linton, Xiaolu Zhao

The problem of jump detection is studied for high-frequency data using a price duration approach. We propose a novel estimator that separates both the contribution of microstructure noise and that of "large" price jumps from the price process, which may have interesting implications on asset pricing and forecasting problems. We show the asymptotic normality of our estimator and suggests practical guidelines for determining the tuning parameter thereof. Making a comparison with the "star performers" in a recent comprehensive review, we show that our method performs well via extensive simulation studies.

CO694 Room R04 CRYPTOCURRENCY ANALYTICS Chair: Marcell Tamas Kurbucz

C0334: **No cryptocurrency experience required: Managerial characteristics in cryptocurrency fund performance**
Presenter: **Andrew Urquhart**, ICMA Centre, Henley Business School, University of Reading, United Kingdom
Co-authors: Pengfei Wang, Yii Li

The aim is to investigate the managerial characteristics, which are determinants of cryptocurrency fund performance. We compile a unique dataset of cryptocurrency fund performance and characteristics of the funds and their managers. We document substantial differences in cryptocurrency fund manager ability in terms of monthly excess returns as well as risk-adjusted returns. In particular, we find that managers with a PhD or MBA tend to have significantly higher excess returns while PhD managers also generate significantly higher risk-adjusted returns. Further, the results also show that managers with previous hedge fund experience generate significantly higher appraisal ratios indicating their investment-picking ability obtained from their previous experiences. However, we also find that cryptocurrency experience offers no explanatory power indicating that trading cryptocurrencies successfully does not require any specific knowledge. Overall, our findings are consistent with the conventional wisdom that manager qualifications and experience play a significant role in fund performance.

C0463: **Empirical comparison of preferential attachment and linking statistics in Bitcoin and Ethereum**
Presenter: **Daniel Kondor**, SMART, Singapore
Co-authors: Gabor Vattay, Istvan Csabai, Jozsef Steger, Nikola Bulatovic

Cryptocurrencies have presented a disruptive change for both economics and computer science. Considering the list of transactions as an evolving network, cryptocurrencies are among the largest real-world networks that can be analyzed by the scientific community, with several hundred million total edges. While there is significant interest in how cryptocurrencies work from a network science perspective, we still do not have a comprehensive understanding of which are the relevant processes that shape the network structure. We evaluate key network characteristics on the Bitcoin and Ethereum transaction networks, the two most popular cryptocurrencies. We specifically look at network evolution and the dynamics of how nodes gain new transaction partners and gain or lose balance. We show that a process of preferential attachment continues to be determinant for both cryptocurrencies and is robust concerning the time period analysed and the method used to reconstruct the transaction network. During our analysis, we perform an in-depth comparison among Bitcoin and Ethereum, focusing on comparing the transaction dynamics of regular addresses in the two systems and between addresses and smart contracts in Ethereum. In all cases, we evaluate correlations between node degree, activity and wealth.

C0622: **Frequency decomposition of crypto connectedness**
Presenter: **Jan Sila**, Univerzita Karlova, Czech Republic
Co-authors: Ladislav Kristoufek

The aim is to describe the connectedness of cryptocurrency markets that arise with different responses to shocks. Due to a framework introduced previously, we can detect different responses in the time-frequency domain. As market agents operate on different investment or speculative

horizons, spectral representation of variance decomposition allows us to reveal such information. As cryptocurrencies are considered to be governed by algorithmic trading, we can compare such dynamics on high-frequency or daily data, thus measuring these differences at arbitrary frequencies. We then compare the dynamics with traditional financial assets.

CC814 Room R06 CONTRIBUTIONS IN APPLIED ECONOMETRICS II Chair: Pilar Poncela

C1081: **Variable selection in text regression: The case of short texts**
Presenter: **Marzia Freo**, University of Bologna, Italy
Co-authors: Alessandra Luati
Communication through websites is often fast and characterised by short texts, made of few words, such as titles, image captions, questions and answers and tweets or posts in social media. The class of supervised learning methods for the analysis of short texts is explored. The aim is to assess the effectiveness of text data in social sciences when they are used as explanatory variables in parametric regression models. We compare the results obtained by several variants of the lasso, screening-based methods and randomization-based models, such as sure independent screening and stability selection. A widely applied unsupervised learning method for topic modelling, Latent Dirichlet Allocation, is also considered. The perspective is primarily empirical, and our starting point is the analysis of two real datasets, though bootstrap replications of each dataset. The first case study aims at explaining price variations based on the information contained in the description of items on sale on e-commerce platforms by using the internet as a source of data. The second case study is concerned with the informative content if open questions inserted in a questionnaire on overall satisfaction ratings. The two case studies are different in nature and thus representative of different kinds of short texts, as, in the first application, a short descriptive and objective text is considered, whereas, in the second case study, the short text is subjective and emotional.

C0558: **What drives labour market success: Empirical analysis of university graduates**
Presenter: **Sylwia Roszkowska**, University of Lodz, Poland
Co-authors: Paulina Hojda, Mariusz Trojak
The purpose is to assess the factors determining the success of university graduates in the labour market. Success is measured in a narrow way as the simultaneous occurrence of work in accordance with the field of study, pay above the average in the economy and job satisfaction. We also use the measure of broad success - in this case, we talk about success when at least one of the above mentioned features exists. We use data from the survey of graduates of the oldest university in Poland. We analyse four editions of this study. The obtained results indicate that success on the labour market is influenced not only by the completed field of study, but also by the activity undertaken during the studies, type of studies and demographic features. The results also confirm the premises of behavioural theory.

C0647: **A pound centric look at the pound vs krona exchange rate movement from 1844 to 1965**
Presenter: **Andrew Clark**, University of Reading, United Kingdom
A longitudinal (1844-1965) study of the Pound Krona exchange rate is conducted utilizing London Times article news sentiment, gold price, GDP, and other relevant metrics to create a dynamic systems state-based model to predict the Pound Krona yearly exchange rate. The created model slightly outperforms a naive random walk forecasting model.

CG747 Room R07 CONTRIBUTIONS IN EMPIRICAL MACROECONOMICS Chair: Luis Filipe Martins

C0382: **Bremia: A study of the impact of Brexit and COVID-19 based on bond prices**
Presenter: **Corrado Macchiarelli**, National Institute of Economic and Social Research, United Kingdom
Co-authors: Jagjit Chadha, Arno Hantzsche, Cyrille Lenoel, Sathya Mellina
Many financial prices reacted violently to the result of the UK's advisory referendum held on 23 June 2016 and to the spread of COVID-19. Subsequently, financial prices have proved significantly less volatile, both unconditionally and in response to the news. We want to understand what sovereign bond prices might have been telling us about the likely state of the British economy under an exit from the European Union and its re-orientation in light of COVID-19. To do so, we model the factors determining the term structure of interest rates and find that bond yields are driven by macroeconomic factors, as well as by central bank communication which we quantify using text mining techniques. When we map our results to movements in response to the news, we find that bond yields decline in anticipation of more expansionary monetary policy. A plausible explanation is that bond markets have started to anticipate further QE. In this sense, the change in expectations about monetary policy appears to have offset any Brexit-related rise in any primitive risk premium. Instead, our preliminary results for COVID-19 are rather mixed, arguably because of a lack of data-points, and we treat them as very tentative. We find that COVID-19 uncertainty did not significantly increase term premia at the 10-year maturity in the same manner as Brexit-related uncertainty. This seems to be consistent with the view that market participants expect a hefty but not enduring crisis.

C0940: **Inequality and growth in France: A wavelet analysis**
Presenter: **Alessandro Pietropaoli**, Cote dÁzur University, France
The relationship between inequality and growth is a traditional, but still unsolved puzzle in economics. In particular, the time horizon at which the relation is investigated plays a crucial role in shaping the empirical results. We focus on the case of France, for which historical data (1915-2016) on several inequality measures, GDP growth rates and a set of important covariates are now accessible. The availability of time series that go sufficiently far back allows us to exploit continuous wavelet tools to shed light on the time scale relation between income distribution and growth. By performing spectral analysis as a function of time, wavelet techniques are very well-suited for examining time-varying relationships across frequencies. We show that the relationship is particularly strong in the medium and in the long term, while weaker and quite unstable in the short run. Further, the lead/lag relationship cannot be taken for granted, since the leading variable tends to change over time and across frequencies. Finally, in the long run, when the association is particularly strong and significant, we find that the impact of unequal income distribution on growth turns out to be negative.

C1033: **Estimating financial frictions under learning**
Presenter: **Jacek Suda**, FAME|GRAPE, Poland
Co-authors: Patrick Pintus, Burak Turgut
The collapse of the housing market in the second half of the 2000s triggered the credit crisis and impelled the U.S. economy into the Great Recession. The crisis underscored the importance of both expectations and the role that financial markets (and the housing market in particular) played in the economy and brought back the interest in understanding the links between these markets and the macroeconomy. We study the quantitative implications of the departure from the rational expectations for the financial crisis. We introduce constant-gain adaptive learning into a medium-scale DSGE model with credit-constrained agents. Using data on leverage and usual macroeconomic variables for the period 1975-2008, we estimate the model both under rational expectations and adaptive learning using Bayesian techniques. We find that the learning model has a better fit than its rational expectations version. We find that in the model with learning large negative collateral shock observed in the 2008Q3 had a significant effect, which is compounded by the imperfect information about this process. We assess whether alternative monetary policy rules of

the central bank could prevent the housing market bubble from arising and in the aftermath of the crisis. The preliminary results suggest that neither average inflation targeting nor price level targeting would not have been effective in avoiding the crisis, but they do affect the recovery period.

Sunday 20.12.2020 13:15 - 14:55 Parallel Session I – CFE-CMStatistics

EO696 Room R11 INFERENCE FOR FUNCTIONAL PARAMETERS Chair: Dominik Liebl

E0293: **Feature extraction for functional time series: Theory and application to NIR spectroscopy data**
Presenter: **Yang Yang**, Monash University, Australia
Co-authors: Han Lin Shang, Yanrong Yang

A novel method is proposed to extract global and local features of functional time series. The global features concerning the dominant modes of variation over the entire function domain, and the local features of function variations over particular short intervals within the function domain, are both important in functional data analysis. Functional principal component analysis (FPCA), though a key feature extraction tool, only focuses on capturing the dominant global features, neglecting highly localized features. We introduce an FPCA-BTW method that initially extracts global features of functional data via FPCA, and then extracts local features by block thresholding of wavelet (BTW) coefficients. Using Monte Carlo simulations, along with an empirical application on near-infrared spectroscopy data of wood panels, we illustrate that the proposed method outperforms competing methods, including FPCA and sparse FPCAin the estimation functional processes. Moreover, extracted local features inheriting serial dependence of the original functional time series contribute to more accurate forecasts. Finally, we develop asymptotic properties of FPCA-BTW estimators, discovering the interaction between convergence rates of global and local features.

E0680: **A dynamic factor model for functional time series: Identification, estimation, and prediction**
Presenter: **Sven Otto**, University of Bonn, Germany
Co-authors: Nazarii Salish

A fully functional factor model is proposed in which both the common component and the idiosyncratic component are random elements of the Hilbert space of L2 integrable functions on a bounded domain. We assume that the factors are dynamic and follow a vector autoregressive process, while the errors are H-white noise. Together with suitable conditions for the factors and loading functions, the common factor dynamics allows us to identify the factor and the error component separately. By applying the least-squares principle, we obtain an estimator based on functional principal components which are shown to be consistent. The number of factors and lags are jointly estimated by a forecast error based information criterion. For curve predictions, we suggest the minimum mean square error forecast from the dynamic functional factor model, and prediction bands are provided under additional distributional assumptions. Finally, the methodology is applied to the problem of yield curve modeling and forecasting. In an out-of-sample experiment, it is shown that the predictions can be significantly improved when compared to the predictor from the dynamic Nelson-Siegel model, which is the most commonly used term structure model for yield curves.

E0421: **Functional delta residuals and applications to functional effect sizes**
Presenter: **Fabian Telschow**, University of California San Diego, United States
Co-authors: Samuel J Davenport, Armin Schwartzman

Given a functional central limit (fCLT) and a parameter transformation, we use the functional delta method to construct random processes, called functional delta residuals, which asymptotically have the same covariance structure as the transformed limit process. As motivation for this methodology, we provide the formal application of these residuals to a functional version of the effect size parameter Cohen's d. We prove a multiplier bootstrap fCLT theorem for these transformed residuals and show how this can be used to construct simultaneous confidence bands (SCBs) for Cohen's d. The performance and necessity of such residuals are illustrated in a simulation experiment for the covering rate of SCBs for the functional Cohen's d parameter and an application to CoPE sets in brain imaging is presented.

E0716: **Simultaneous inference in functional data analysis**
Presenter: **David Degras**, University of Massachusetts Boston, United States

Statistical methods will be discussed for simultaneously inferring functional parameters such as mean, covariance, and regression functions. From a theoretical perspective, we will start with the standard case of i.i.d., densely observed functional data (FD) and then discuss challenges posed by other observation schemes (e.g. sparse or partially observed FD) and dependence mechanisms (e.g. functional time series, spatial FD). A numerical study will provide insights into the computational and statistical performances of simultaneous confidence bands methods. Finally, we will consider recent lines of FD research, such as manifolds and partially observed FD.

EO604 Room R12 STATISTICAL PROBLEMS UNDER PRIVACY CONSTRAINTS Chair: Cristina Butucea

E0182: **Minimax optimal goodness-of-fit testing under local differential privacy**
Presenter: **Joseph Lam-Weil**, Magdeburg University, Germany
Co-authors: Jean-Michel Loubes, Beatrice Laurent-Bonneau

The consequences of local differential privacy constraints on goodness-of-fit testing are considered, i.e. the statistical problem assessing whether sample points are generated from a fixed density or not. The observations are hidden and replaced by a stochastic transformation satisfying the local differential privacy constraint. The focus will be on the lower bound, leading to the minimax optimality of our result over Besov balls.

E0428: **Locally private non-asymptotic testing of discrete distributions is faster using interactive mechanisms**
Presenter: **Thomas Berrett**, University of Warwick, United Kingdom
Co-authors: Cristina Butucea

Recent work on hypothesis testing under local differential privacy is presented. We find separation rates for testing multinomial or more general discrete distributions under the LDP constraint. The upper bounds are established by constructing efficient randomized algorithms and test procedures, in both the case where only non-interactive privacy mechanisms are allowed and also in the case where all sequentially interactive privacy mechanisms are allowed. We find that the separation rates are faster in the latter case. The lower bounds are based on general information theoretical bounds that allow us to establish the optimality of our algorithms among all pairs of privacy mechanisms and test procedures, in most usual cases. Considered examples include testing uniform, polynomially and exponentially decreasing distributions.

E1137: **Estimating linear and quadratic functionals under local differential privacy**
Presenter: **Lukas Steinberger**, University of Vienna, Austria

In the local paradigm of differential privacy, we study the problem of estimating a functional of the true unknown data generating distribution. For linear functionals over convex parameter spaces, we develop a general minimax theory of private estimation. We find that appropriately privatized sample mean type estimators are always minimax optimal and, in particular, that local privacy protocols that allow for some sequential interaction between data providers do not improve significantly over purely non-interactive protocols. These facts are no longer true for non-linear functionals. For estimation of the integrated square of the density over Besov spaces, we present a locally private estimator that is sequential in nature and improves over the best non-interactive procedure in terms of minimax rate.

E1105: **Differentially private sub-Gaussian location estimators**
Presenter: **Victor-Emmanuel Brunel**, ENSAE ParisTech, France
Co-authors: Marco Avella-Medina

The focus is on the problem of estimating a location parameter with differential privacy guarantees and sub-Gaussian deviations. Namely, we

propose estimators that achieve an error with sub-Gaussian tails and satisfy the standard differential privacy constraint, even when the data only have a few finite moments. Moreover, our method does not require the unknown location parameter to be bounded in a known region, unlike previous results.

EO283 Room R13 RECENT ADVANCES IN EXTREME VALUE ANALYSIS Chair: Gilles Stupfler

E0290: **Extremile regression**
Presenter: **Abdelaati Daouia**, Fondation Jean-Jacques Laffont, France
Co-authors: Irene Gijbels, Gilles Stupfler

Regression extremiles define a least squares analogue of regression quantiles. They are determined by weighted expectations rather than tail probabilities. Of special interest is their intuitive meaning in terms of expected minima and maxima. Their use appears naturally in risk management where, in contrast to quantiles, they fulfill the coherency axiom and take the severity of tail losses into account. In addition, they are co-monotonically additive and belong to both families of spectral risk measures and concave distortion risk measures. The aim is to provide the first detailed study exploring implications of the extremile terminology in a general setting of presence of covariates. We rely on local linear (least squares) check function minimization for estimating conditional extremiles and deriving the asymptotic normality of their estimators. We also extend extremile regression far into the tails of heavy-tailed distributions. Extrapolated estimators are constructed and their asymptotic theory is developed. Some applications to real data are provided.

E0438: **Joint inference on extreme expectiles for multivariate heavy-tailed distributions**
Presenter: **Simone Padoan**, Bocconi University, Italy
Co-authors: Gilles Stupfler

The notion of expectiles, originally introduced in the context of testing for homoscedasticity and conditional symmetry of the error distribution in linear regression, induces a law-invariant, coherent and elicitable risk measure that has received a significant amount of attention in actuarial and financial risk management contexts. Several recent papers have focused on the behaviour and estimation of extreme expectile-based risk measures and their potential for risk management. Joint inference of several extreme expectiles has however been left untouched; in fact, even the inference of a marginal extreme expectile turns out to be a difficult problem in finite samples. We investigate the simultaneous estimation of several extreme marginal expectiles of a random vector with heavy-tailed marginal distributions. This is done in a general extremal dependence model where the emphasis is on pairwise dependence between the margins. We use our results to derive accurate confidence regions for extreme expectiles, as well as a test for the equality of several extreme expectiles. Our methods are showcased in a finite-sample simulation study and on real financial data.

E0641: **Records for time-dependent stationary Gaussian sequences**
Presenter: **Amir Khorrami Chokami**, University of Turin, Italy
Co-authors: Michael Falk, Simone Padoan

For a zero-mean, unit-variance stationary univariate Gaussian process we derive the probability that a record at the time n, say X_n, takes place, and derive its distribution function. We study the joint distribution of the arrival time process of records and the distribution of the increments between records. We compute the expected number of records. We also consider two consecutive and non-consecutive records, one at time j and one at time n, and we derive the probability that the joint records (X_j, X_n) occur, as well as their distribution function. The probability that the records X_n and (X_j, X_n) take place and the arrival time of the n-th record are independent of the marginal distribution function, provided that it is continuous. These results actually hold for a strictly stationary process with Gaussian copulas.

E1015: **On second-order automatic bias reduction for extreme expectile estimation**
Presenter: **Antoine Usseglio-Carleve**, Inria, France
Co-authors: Stephane Girard, Gilles Stupfler

Expectiles induce a law-invariant risk measure that has recently gained popularity in actuarial and financial risk management applications. They are determined by tail expectations and induce a coherent and elicitable risk measure, while quantiles are calculated in terms of tail probabilities only and are not coherent, and the quantile-based Expected Shortfall is not elicitable. Expectiles have therefore been suggested as serious candidates for a standard risk measure. Their estimation in the heavy-tailed framework is not without difficulties; currently available estimators of extreme expectiles are typically biased and hence may show poor finite-sample performance even in fairly large samples. We focus here on the construction of bias-reduced extreme expectile estimators for heavy-tailed distributions. The rationale for our construction hinges on a careful investigation of the asymptotic proportionality relationship between extreme expectiles and their quantile counterparts, as well as of the extrapolation formula motivated by the heavy-tailed context. We accurately quantify and estimate the bias incurred by the use of these relationships when constructing extreme expectile estimators. This motivates the introduction of a class of bias-reduced estimators whose asymptotic properties are rigorously shown, and whose finite-sample properties are assessed on a simulation study and several samples of real data.

EO478 Room R14 RECENT ADVANCES FOR THE MODELLING OF COMPLEX DATA Chair: Ines Varas

E0849: **Grid-uniform copulas and rectangle exchanges: Model and Bayesian inference for a rich class of copula functions**
Presenter: **Nicolas Kuschinski**, Pontificia Universidad Catolica de Chile, Chile
Co-authors: Alejandro Jara

Copula-based models provide a great deal of flexibility in modelling multivariate distributions, allowing for the specifications of models for the marginal distributions separately from the dependence structure (copula) that links them to form a joint distribution. Choosing a class of copula models is not a trivial task, but it can be simplified by relying on rich classes of copula functions. We introduce a novel class of grid-uniform copula functions, which is dense (in the Hellinger sense) in the space of all continuous copula functions. We propose a Bayesian model based on this class and develop an efficient Markov chain Monte Carlo algorithm for exploring the corresponding posterior distribution in arbitrarily many dimensions, allowing for the semiparametric or nonparametric modelling of continuous joint distributions.

E0860: **Spatial modeling of georeferenced count data**
Presenter: **Diego Morales Navarrete**, Pontificia Universidad Catolica de Chile, Chile
Co-authors: Moreno Bevilacqua, Luis Mauricio Castro Cepero, Christian Caamano Carrillo

Modeling spatial data is a challenging task in statistics. In many applications, the observed data can be modeled using Gaussian, skew-Gaussian, or even restricted random field models. However, in several fields, such as population genetics, epidemiology, and aquaculture, the data of interest are counts, and therefore the mentioned models are not suitable for their analysis. Consequently, there is a need for spatial models that can adequately describe data coming from counting processes. Three approaches are used to model this type of data: GLMMs with Gaussian random field (GRF) effects, hierarchical models, and copula models. Unfortunately, these approaches do not explicitly characterize the random field like their q-dimensional distribution or correlation function. It is important to stress that GLMMs and hierarchical models induce a discontinuity in the path. Hence, we propose a novel approach to efficiently and accurately model spatial count data to deal with this. Briefly, starting from independent copies of a parent GRF, a set of transformations can be applied, and the result is a non-Gaussian random field. This approach is based on a random field characterization for count data that inherit some well-known geometric properties from GFRs.

E1118: **An extension of multivariate models for a monotone disease under misspecification**
Presenter: **Ines Varas**, Pontificia Universidad Catolica de Chile, Chile

Caries experience (CE) is an essential topic in dental and medical sciences. This process is prone to misclassification because of several reasons. The CE process is a monotone disease because when a tooth with caries has been found, it can reverse its state. Hidden Markov models (HHMs) have been applied in several research areas to connect latent and observed states of a process along time. These models allow the definition of different types of misclassification structures. Most of the research already developed in this framework considers a dichotomic definition of both the latent and observed processes. Nevertheless, the definition of two stages processes could be restrictive for modelling the CE process in the sense that the model does not describe the intermediate stages of it. We propose a hidden Markov model for a misclassified longitudinal monotone process defined in more than two categories to evaluate the prevalence and incidence of the process. Identifiability restrictions to obtain consistent and interpretable parameters are evaluated.

E1168: **On the support of some random contiguous set partitions for change point analysis**
Presenter: **Jose Javier Quinlan Binelli**, Pontificia Universidad Catolica de Chile, Chile
Co-authors: Alejandro Jara, Mauricio Castro

In Bayesian change point analysis for univariate time series, prior distributions on the space of contiguous set partitions play a key role for change point detection. In this context, mixtures of such distributions are appealing candidates in terms of diversity in the specification of prior beliefs. However, how flexible are these prior processes to describe the number and locations of possible change-points? We will address the previous question in terms of weak support.

EO482 Room R15 RECENT ADVANCES IN MULTIVARIATE AND MULTI-VARIABLE ANALYSIS Chair: Frank Konietschke

E0322: **Multivariate mixed response model with pairwise composite-likelihood method**
Presenter: **Xin Gao**, York University, Canada

In clinical research, study outcomes usually consist of various patients' information corresponding to the treatment. To have a better understanding of the effects of different treatments, one often needs to analyze multiple clinical outcomes simultaneously. At the same time, the data are usually mixed with both continuous and discrete variables. We propose the multivariate mixed response model to implement statistical inference based on the conditional grouped continuous model through a pairwise composite-likelihood approach. We demonstrate the validity and statistical power of the multivariate mixed response model through simulation studies and clinical applications. This composite-likelihood method is advantageous for statistical inference on correlated multivariate mixed outcomes.

E0420: **Ranking procedures for repeated measures design with missing data**
Presenter: **Kerstin Rubarth**, Institute of Biometry and Clinical Epidemiology - Charite Berlin, Germany
Co-authors: Frank Konietschke

A frequently used design in health, medical and biomedical research is the repeated measures design, where units, e.g. patients, are observed several times under different conditions. However, if data is highly skewed, ordinal or even ordered categorical or if sample sizes are small, only a few methods are applicable as many methods, e.g. linear mixed models, rely on the assumption of multivariate normality and a specific covariance matrix structure of the error terms. Additionally, in many studies with repeated measures, missing values are almost certain to occur. Therefore, we propose a purely nonparametric method for the analysis of repeated measures with missing data. The hypotheses are formulated in terms of the nonparametric treatment effect. Global testing and a multiple contrast test procedure, as well as simultaneous confidence intervals, are developed for this design. We present simulation studies, which indicate a good performance of this procedure in terms of the type-I-error rate and the power under different alternatives with a missing rate up to 30%, also under non-normality. A real data example illustrates the application of the newly developed methodology.

E0828: **Multi-category individualized treatment regime**
Presenter: **Jin Xu**, East China Normal University, China
Co-authors: Xinyang Huang

Individualized treatment regimes (ITRs) aim to recommend treatments based on patient-specific characteristics in order to maximize the expected clinical outcome. Outcome weighted learning approaches have been proposed for this optimization problem with a primary focus on the binary treatment case. We propose a general framework for multi-category ITRs using generic surrogate risk. The proposed method accommodates the situations when the outcome takes a negative value and/or when the propensity score is unknown. At the same time, risks caused by different adverse events cannot be ignored. We thus propose another method to estimate an optimal individualized treatment rule that maximizes clinical benefit outcome while having the risk-controlled at the desired level. The proposed procedure employs a novel surrogate loss and a relaxed difference of convex functions algorithm to solve the nonconvex constrained optimization problem. Simulations and real data examples are used to demonstrate the finite sample performance.

E0955: **Weighted empirical and Euclidean likelihood covariate adjustment**
Presenter: **Mihai Giurcanu**, University of Chicago, United States
Co-authors: George Luta, Gary Koch, Kirstine Amris, Pranab Sen

Covariate adjustment is often used in the statistical analysis of randomized experiments to increase the efficiency of estimators of treatment effects. We study covariate adjustment based on the empirical and Euclidean likelihoods and propose weighted versions that arise as natural alternatives. The weighted methods incorporate the auxiliary information that the covariates have equal means among the treatment groups due to randomization. We show that the empirical and the Euclidean likelihoods and their weighted versions are the first-order equivalents to Koch's nonparametric covariance adjustment. Allowing the weights to be negative, the resulting pseudo-Euclidean likelihood is equivalent to Koch's method, and its weighted version can be viewed as a weighted version of Koch's method. In a simulation study, we assess the finite sample properties of the proposed methods. The analysis of a randomized clinical trial data set illustrates an application of these methods to a practical situation.

EO381 Room R16 ADVANCES IN CAUSAL INFERENCE Chair: Daniel Malinsky

E0582: **High-dimensional model-assisted inference for local average treatment effects with instrumental variables**
Presenter: **BaoLuo Sun**, National University of Singapore, Singapore
Co-authors: Zhiqiang Tan

Consider the problem of estimating the local average treatment effect with an instrument variable, where the instrument unconfoundedness holds after adjusting for a set of measured covariates. Several unknown functions of the covariates need to be estimated through regression models, such as instrument propensity score and treatment and outcome regression models. We develop a computationally tractable method in high-dimensional settings where the numbers of regression terms are close to or larger than the sample size. Our method exploits regularized calibrated estimation, which involves Lasso penalties but carefully chosen loss functions for estimating coefficient vectors in these regression models, and then employs a doubly robust estimator for the treatment parameter through augmented inverse probability weighting. We provide rigorous theoretical analysis to show that the resulting Wald confidence intervals are valid for the treatment parameter under suitable sparsity conditions if the instrument propensity score model is correctly specified. Still, the treatment and outcome regression models may be misspecified. For existing high-dimensional methods,

valid confidence intervals are obtained for the treatment parameter if all three models are correctly specified. The proposed methods are evaluated through simulation studies and an empirical application to estimate the returns to education.

E0601: **Semiparametric inference for causal effects in graphical models with hidden variables**
Presenter: **Razieh Nabi**, Johns Hopkins University, United States
Co-authors: Rohit Bhattacharya, Ilya Shpitser
The last decade witnessed the development of algorithms that completely solve the identifiability problem for causal effects in causal graphical models with hidden variables. However, much of this machinery remains underutilized in practice, owing to the complexity of estimating identifying functionals yielded by these algorithms. We describe simple graphical criteria and semiparametric estimators that bridge the gap between identification and estimation for causal effects of a single treatment on an outcome. We first discuss influence function-based doubly robust estimators that cover a significant subset of hidden variable causal models where the effect is identifiable. This allows us to incorporate flexible machine learning methods into causal inference pipelines that go beyond the standard, but often unreasonable, assumption of conditional ignorability. We further characterize an important subset of this class for which we demonstrate how to derive the estimator with the lowest asymptotic variance, i.e., one that achieves the semiparametric efficiency bound. Finally, we consider semiparametric estimators that resemble reweighed influence function-based estimators for any identified single treatment causal effect parameter. A Python software package named Ananke that implements these methods is available.

E0668: **Doubly debiased lasso: High-dimensional inference under hidden confounding**
Presenter: **Zijian Guo**, Rutgers University, United States
Co-authors: Domagoj Cevid
Inferring causal relationships or related associations from observational data can be invalidated by the existence of hidden confounding. We focus on a high-dimensional linear regression setting, where the measured covariates are affected by hidden confounding and propose the *Doubly Debiased Lasso* estimator for individual components of the regression coefficient vector. The advocated method simultaneously corrects both the bias due to the estimation of high-dimensional parameters, as well as the bias caused by the hidden confounding. We establish its asymptotic normality and also prove that it is efficient in the Gauss-Markov sense. The validity of this methodology relies on a dense confounding assumption, i.e. that every confounding variable affects many covariates. The finite sample performance is illustrated with an extensive simulation study and a genomic application.

E0312: **Isotonic regression discontinuity designs**
Presenter: **Andrii Babii**, University of North Carolina, United States
Estimation and inference for the isotonic regression at the boundary of its support are studied. This object is particularly interesting and required in the analysis of monotone regression discontinuity designs. We show that the isotonic regression is inconsistent at the boundary and that consistency can be restored with a suitable boundary correction. The one-sided Brownian motion drives the large sample distribution at the boundary. Since the distribution is not pivotal, we also introduce the trimmed wild bootstrap and show its consistency without subsampling or additional nonparametric smoothing. The results are illustrated for sharp and fuzzy monotone regression discontinuity designs. We find in Monte Carlo experiments that shape restrictions can improve the finite-sample performance of unrestricted estimators dramatically. An empirical analysis of the incumbency effect in the U.S. House elections is provided.

EO468 Room R17 RECENT DEVELOPMENT IN STATISTICAL ANALYSIS OF NETWORK DATA Chair: Binyan Jiang

E0349: **Mentorship and prizewinning in scientific careers**
Presenter: **Yifang Ma**, Southern University of Science and Technology, China
Recent work in computational social science and data science will be reported. Mentorship is arguably a scientist's most significant collaborative relationship, and scientific prizes confer credibility to persons and ideas, provide financial incentives and promote community-building celebrations, which are important to a scientific career. Using new large-scale data from the genealogical and performance records of 10s of thousands of scientists and 3K scientific prizes worldwide, we focus on understanding how the knowledge linkages among prizes and scientists propensities for prizewinning and the link between mentorship and protege success. It is shown that not only the scientists' merit of success increase the probability of prizewinning but also the collaboration network and the genealogical network among scientists play important roles in prizewinning. We also found that mentorship is associated with diverse forms of protege success, significantly increasing proteges' chances of producing celebrated research, being inducted into the National Academy of Science, and achieving superstardom. This can help us to understand the role of mentorship and prizewinning, with strong implications in science development.

E0506: **Higher-order accurate inference for network moments**
Presenter: **Yuan Zhang**, Ohio State University, United States
The aim is to derive and use high-order expansions of network moment statistics for exchangeable network models, including the popular stochastic block model for one- and two-sample network inferences, under very mild assumptions. By this approach, we can achieve the following two goals simultaneously (i) higher-order control of the type I error; and (ii) rate-optimal separation condition on the alternative hypothesis for the test to be consistent. Notice that goal (i) was previously only achieved by computationally expensive bootstrap methods with no power guarantees. We also demonstrate our approach's effectiveness in numerical examples.

E1133: **Change-point estimation in a dynamic stochastic block model**
Presenter: **Monika Bhattacharjee**, IIT BOMBAY, India
Co-authors: Moulinath Banerjee, George Michailidis
The purpose is to provide an extensive investigation of change-point analysis for networks generated by stochastic block models, to identify key conditions for the consistent estimation of the change-point, and to propose a computationally fast algorithm that solves the problem in many settings that occur in applications. We establish rates of convergence and derive the asymptotic distributions of the change point estimators. The results are illustrated on synthetic data. Finally, it discusses challenges posed by employing clustering algorithms in this problem, that require additional investigation for their full resolution.

E0982: **Autoregressive networks**
Presenter: **Binyan Jiang**, The Hong Kong Polytechnic University, Hong Kong
A first-order autoregressive model is proposed for dynamic network processes in which edges change over time while nodes remain unchanged. The model depicts the dynamic changes explicitly. It also facilitates simple and efficient statistical inference such as the maximum likelihood estimators which are proved to be (uniformly) consistent and asymptotically normal. The model diagnostic checking can be carried out easily using a permutation test. The proposed model can apply to any Erdös-Renyi network processes with various underlying structures. As an illustration, an autoregressive stochastic block model has been investigated in depth, which characterizes the latent communities by the transition probabilities over time. This leads to a more effective spectral clustering algorithm for identifyingthe latent communities. Inference for change-points is incorporated into the autoregressive stochastic block model to cater for possible structure changes. The developed asymptotic theory as well as the simulation study affirm the performance of the proposed methods. Application with three real data sets illustrates both relevance and usefulnessof the proposed models.

EO530 Room R18 METHODOLOGICAL AND COMPUTATIONAL ADVANCES IN COPULA MODELS Chair: Elif Acar

E0197: **Quasi-Monte Carlo for multivariate distributions via generative neural networks**
Presenter: **Marius Hofert**, University of Waterloo, Canada
Co-authors: Mu Zhu, Avinash Prasad

A novel approach based on generative neural networks is introduced for constructing quasi-random number generators for multivariate models with any underlying copula in order to estimate expectations with variance reduction. So far, quasi-random number generators for multivariate distributions required a careful design, exploiting specific properties (such as conditional distributions) of the implied copula or the underlying quasi-Monte Carlo point set, and were only tractable for a small number of models. Utilizing specific generative neural networks allows one to construct quasi-random number generators for a much larger variety of multivariate distributions without such restrictions. Once trained with a pseudo-random sample, these neural networks only require a multivariate standard uniform randomized quasi-Monte Carlo point set as input and are thus fast in estimating expectations under dependence with variance reduction. Reproducible numerical examples are considered to demonstrate the approach. Emphasis is put on ideas rather than mathematical proofs.

E0323: **Proxy variables to common factors and parameter estimation in factor copula models**
Presenter: **Pavel Krupskiy**, Melbourne University, Australia
Co-authors: Harry Joe

Factor copula models assume observed variables are independent conditional on one or several unobserved factors. We show that, under some mild assumptions, proxy variables to the unobserved factors can be obtained from the observed variables when the dimension is large. These proxy variables can help to select appropriate linking copulas in some factor copula models and to perform numerically faster maximum likelihood estimation of parameters of these high-dimensional copulas. A simulation study shows that parameter estimates obtained using the proxy variable approach are close to those obtained using the maximum likelihood approach. The proxy variable approach is used to analyze a financial data set of stock returns from different sectors.

E0292: **Time-heterogeneous D-vine copula model for longitudinal data**
Presenter: **Md Erfanul Hoque**, University of Manitoba, Canada
Co-authors: Elif Acar, Mahmoud Torabi

Longitudinal studies collect repeated measurements from subjects over time to understand the dependence mechanisms among these measurements. However, in many cases, the number and timing of measurements differ across study subjects so that the data may be unbalanced and unequally spaced and may have an impact on the dependence structure of such data. Hence, statistical analysis of such data requires accounting for both the unbalanced study design and the spacing of repeated measurements. We propose a time-heterogeneous D-vine copula model that allows for time adjustment in the dependence structure of unequally spaced and potentially unbalanced longitudinal data. Moreover, we investigate the asymptotic properties of the parameter estimates under proposed models. The performance of the time-heterogeneous D-vine copula models is evaluated through simulation studies and by a real data application.

EO752 Room R20 ANALYZING AND PREDICTING DIFFERENT OUTCOMES IN CLINICAL STUDIES Chair: Eleni-Rosalina Andrinopoulou

E0202: **Minimizing the burden of invasive procedures via personalized scheduling**
Presenter: **Dimitris Rizopoulos**, Erasmus University Rotterdam, Netherlands

In early-stage chronic diseases, invasive procedures, such as biopsies are used for diagnosing disease progression. Patients typically undergo these invasive tests in a fixed one-size-fits-all manner. An example of such a setting is prostate cancer patients with low-grade tumors. Namely, patients are closely monitored using blood tests. Still, the decision to treat is based on prostate biopsies. The problem is that biopsies are painful and lead to complications. The current standards are to perform biopsies for all patients every one or three years. We argue that a better approach is to opt for personalized test schedules. Our approach utilizes the progression-risk of each patient. It aims to balance the number of tests (burden) and time delay in detecting progression (shorter is beneficial). Our approach uses a novel statistical modeling framework called joint models for time-to-event and longitudinal data. Using these models, we consolidate patients' longitudinal data (e.g., biomarkers) and previous tests' results into individualized future cumulative-risk of progression. We then create personalized schedules by planning tests on future visits where the predicted cumulative-risk is above a threshold. To find the optimal risk threshold, we minimize a utility function of the expected number of tests and expected time delay in detecting progression. These two quantities are estimated in a patient-specific manner, using a patient's predicted risk profile.

E0218: **Longitudinal modeling of disease progression biomarkers in the latent disease timescale: Example of Alzheimer's disease**
Presenter: **Cecile Proust-Lima**, INSERM, France
Co-authors: Jeremie Lespinasse, Carole Dufouil

Alzheimer's disease and related disorders (ADRD) are characterized by progressive changes in multiple components, including protein accumulation in the brain, brain atrophies, and cognitive dysfunction. Understanding the sequence and timing of such deteriorations is paramount to refine patient stratification and facilitate earlier diagnosis. However, their modeling faces a fundamental statistical challenge: the timescale is not known. Usual timescales are inappropriate: (i) time of clinical diagnosis is not an option as most of the deteriorations appear years to months before, (ii) time since inclusion does not have any biological meaning, and (iii) chronological age induces too much inter-individual heterogeneity as people do not age similarly and ADRD onset may arise at different ages. We discuss how the mixed model theory applied to multivariate longitudinal biomarkers can be used to realign individual trajectories into a common latent disease time while taking into account the specificities of the biomarkers. We then illustrate the method to describe the sequence of progression of 12 biomarkers in the French clinic-based Memento study. Beyond the sequence of biomarker degradation, this methodology may evaluate at what stage of the disease an individual is by providing a prediction of his/her individual disease time. This has the potential for earlier diagnosis.

E0887: **Estimating the effect of insulin use on outcomes in people with cystic fibrosis-related diabetes using causal prediction**
Presenter: **Emily Granger**, London School of Hygiene and Tropical Medicine, United Kingdom
Co-authors: Freddy Frost, Ruth Keogh

Cystic fibrosis-related diabetes (CFRD) is associated with poor clinical outcomes for people with cystic fibrosis. Insulin is recommended as the only treatment for CFRD, but little is known about the consequences of long-term insulin use. Given the relatively low numbers of people with CFRD, running sufficiently powered randomised control trials is extremely challenging. We, therefore, aimed to estimate the effect of long-term insulin use on health outcomes in people with CFRD using longitudinal observational data. In a randomised trial, we would compare average outcomes in those randomised to treatment to those randomised to non-treatment. However, with observational data treatment status depends on individual characteristics that also affect the outcome and the treated and untreated are not directly comparable. We use counterfactual prediction models to predict the counterfactual outcome for each individual, i.e., the expected outcome we would have observed if the treated group were untreated and vice-versa. We used data from the UK cystic fibrosis registry. The longitudinal nature of these data gave rise to several challenges, such as time-dependent confounding, time-dependent eligibility and uncertainty in the direction of causal pathways. We will discuss how we tackled these challenges using two different approaches to counterfactual prediction: inverse-probability-of-treatment weighting of marginal structural models and the g-formula.

E0998: **Marginal structural models with joint exposure to assess variations to chemotherapy intensity**
Presenter: **Marta Fiocco**, Leiden University, Netherlands

Marginal structural models are causal models designed to adjust for time-dependent confounders in observational studies with dynamically adjusted treatments. They are robust tools to assess causality in complex longitudinal data. A marginal structural model is proposed with an innovative dose-delay joint-exposure model for Inverse Probability of Treatment Weighted estimation of the causal effect of therapy modification. The model is motivated by a clinical question concerning the possibility of reducing dosages in a regimen. It is applied to data from a randomized trial of chemotherapy in osteosarcoma, an aggressive primary bone-tumor. This talk focuses on the clinical dynamical process of adjusting the therapy according to patients toxicity history, and the causal effect on the outcome of interest of such therapy modifications. Depending on patients toxicity levels, variations to therapy intensity may be achieved by physicians through a reduction or a delay of the next planned dose. Therefore, negative feedback is present between exposure to cytotoxic agents and toxicity levels, which acts as time-dependent confounders. The construction of the model is presented, and the high complexity and entanglement of chemotherapy data are discussed. Built to address dosage reductions, the model suggests that delays in therapy administration should be avoided.

EO652 Room R21 RECENT ADVANCES IN SEMIPARAMETRIC SURVIVAL ANALYSIS Chair: Dennis Dobler

E0364: **A presmoothing approach for estimation in mixture cure models**
Presenter: **Eni Musta**, University of Amsterdam, Netherlands
Co-authors: Valentin Patilea, Ingrid Van Keilegom

A challenge when dealing with survival data is accounting for a cure fraction, meaning that some subjects will never experience the event of interest. In this context, mixture cure models have been frequently used to estimate both the probability of being cured and the time to event for the susceptible subjects, by usually assuming a parametric (logistic) form of the incidence and a semiparametric Cox proportional model for the latency. The maximum likelihood estimator, implemented in the R package smcure, is then typically used for estimating the regression parameters and the cumulative hazard. We propose a new estimation procedure which, in the first stage, focuses on direct estimation of the parametric cure probability without using distributional assumptions on the latency. It relies on a preliminary nonparametric estimator for the incidence, which is then projected on the parametric logistic class of functions. In the second stage, the survival distribution of the uncured subjects is estimated by maximizing the Cox component of the likelihood. The estimators are shown to be consistent and asymptotically normally distributed, while simulations suggest that presmoothing often improves parameter estimation for small and moderate sample size. The proposed procedure is applied to two medical datasets about studies of patients with melanoma cancer.

E0616: **Semiparametric likelihood inference for heterogeneous survival data under random double truncation**
Presenter: **Achim Doerre**, University of Rostock, Germany

When survival data are sampled exclusively from already failed units during a fixed data collection period, the lifetimes are subject to random double truncation. We study one important special case of this selective sampling setup, in which units originate from different subgroups with varying lifetime distributions. Point processes are used to model the random emergence of units in the population and sample. We investigate semiparametric likelihood inference to account for selection bias. In addition, we show how information on the pattern of unit emergence can be incorporated to obtain estimators with increased precision. Strategies for alleviating numerical issues and improving performance in the implementation are discussed. We explore the finite-sample properties in a simulation study and demonstrate the suggested approach on a large dataset.

E0987: **Efficient and reliable inference in nested case-control studies**
Presenter: **Dennis Dobler**, Vrije Universiteit Amsterdam, Netherlands
Co-authors: Jan Feifel

Nested case-control designs are of great importance in time-to-event studies with rare outcomes or expensive covariate evaluations. Such designs allow for a (random) selection of only very few controls for each case while, at the same time, not much power is lost compared to the full evaluation. We propose a resampling algorithm for the approximation of the distribution of estimators - like the cumulative hazard - whose complexity grows only linearly in sample size and independently of the number of controls per case. The algorithm works for various sampling designs (counter-matching and more customized designs) and minimal assumptions on the dependence between censoring and event times are required. Theoretical validity with growing sample size is proved rigorously, and simulation results confirm the practical usefulness of our approach.

E0658: **Wild bootstrap confidence bands for the cumulative incidence function in Fine-Gray models**
Presenter: **Marina Tiana Dietrich**, Vrije Universiteit Amsterdam, Netherlands
Co-authors: Dennis Dobler, Mathisca de Gunst

The Fine and Gray model in the competing risks setting with censoring-complete data is considered. The goal is to construct asymptotically valid time-simultaneous confidence bands for the cumulative incidence function using the wild bootstrap. To verify the underlying theory, they use elegant and general martingale arguments. Especially for small samples, the flexibility of the wild bootstrap procedure with possibly non-normal multipliers seems preferable over a Gaussian approximation because the latter is only aiming at the asymptotic distribution. The performance of the wild bootstrap confidence bands for different types of multipliers and weight functions is empirically studied through simulations. In addition, they illustrate the developed method by investigating the impact of Pneumonia for intensive care unit patients on the probabilities of alive discharge competing with hospital death.

EO486 Room R23 FLEXIBLE BAYESIAN MODELS FOR COMPLEX DATA Chair: Alessandra Guglielmi

E0376: **MCMC computations for Bayesian mixture models using repulsive point processes**
Presenter: **Mario Beraha**, Politecnico di Milano, Universita di Bologna, Italy
Co-authors: Raffaele Argiento, Alessandra Guglielmi, Jesper Moeller

Repulsive mixture models have recently gained popularity for Bayesian cluster detection. Compared to more traditional mixture models, there is empirical evidence suggesting that repulsive mixture models produce a smaller number of well-separated clusters. The most commonly used methods for posterior inference either require to fix a priori the number of components or are based on reversible jump MCMC computation. We present a general framework for mixture models, when the prior of the cluster centres is a finite point process depending on a hyperparameter - not only a Poisson or determinantal point process (DPP) as previously considered in the literature but also a repulsive point process specified by a density which depends on an intractable normalizing constant. By investigating the posterior characterization of this class of mixture models, we derive an MCMC algorithm which avoids the well-known difficulties associated with reversible jump MCMC computation. In particular, we use an ancillary variable method, which depends on perfect simulation, to overcome the problem of having a ratio of normalizing constants in the Hastings ratio when making posterior simulations for full conditional of the hyperparameter. In several simulation studies and an application on sociological data, we illustrate the advantage of our new methodology over existing methods, and we compare the use of a DPP or a repulsive Gibbs point process prior model.

E0598: **Clustering and prediction in the presence of variable dimension covariate vectors**
Presenter: **Fernando Quintana**, Pontificia Universidad Catolica de Chile, Chile
Co-authors: Garritt Page, Peter Mueller

In many applied fields incomplete covariate vectors are commonly encountered. It is well known that this can be problematic when making inference on model parameters, but its impact on prediction performance is less understood. We develop a method based on covariate dependent partition models that seamlessly handles missing covariates while completely avoiding any type of imputation. The method we develop allows in-sample predictions as well as out-of-sample prediction, even if the missing pattern in the new subjects' incomplete covariate vector was not seen in the training data. Any data type, including categorical or continuous covariates are permitted. In simulation studies, the proposed method compares favorably.

E0625: **Bayesian factor analysis for estimating a non-randomised intervention via multi-outcomes observational panel data**
Presenter: **Silvia Montagna**, University of Turin, Italy
Co-authors: Pantelis Samartsidis, Daniela de Angelis, Shaun Seaman, Andre Charlett, Matthew Hickman

The problem of estimating the effect of a non-randomized binary intervention on multiple outcomes of interest is addressed by using time series data on units that received the intervention (treated) and units that did not (controls). One popular estimation method in this setting is based on the factor analysis (FA) model within the Rubin causal inference framework. We propose a model that extends the FA model for estimating intervention effects by jointly modelling the multiple outcomes to exploit shared variability, and assuming an auto-regressive structure on factors to account for temporal correlations. Our simulation studies show that the proposed method can improve the precision of the intervention effect estimates and achieve better control of the type I error rate (compared with the FA model), especially when either the number of pre-intervention measurements or the number of control units is small. We apply our method to estimate the effect of stricter alcohol licensing policies on alcohol-related harms.

E0684: **Bayesian nonparametric priors for hidden Markov random fields**
Presenter: **Julyan Arbel**, Inria, France
Co-authors: Florence Forbes, Hongliang Lu

One of the central issues in statistics and machine learning is how to select an adequate model that can automatically adapt its complexity to the data. We focus here on the issue of determining the structure of clustered data, both in terms of finding the appropriate number of clusters and of modelling the right dependence structure between the observations. Bayesian nonparametric (BNP) models, which do not impose an upper limit on the number of clusters, are appropriate to avoid the required guess on the number of clusters but have been mainly developed for independent data. In contrast, Markov random fields (MRF) have been extensively used to model dependencies in a tractable manner but usually reduce to finite cluster numbers when clustering tasks are addressed. The main contribution is to propose a general scheme to design tractable BNP-MRF priors that combine both features: no commitment to an arbitrary number of clusters and dependence modelling. A key ingredient in this construction is the availability of a stick-breaking representation which has the three-fold advantage of 1) extending standard discrete MRFs to infinite state space, 2) designing a tractable estimation algorithm using variational approximation and 3) deriving theoretical properties on the predictive distribution and the number of clusters of the proposed model. This approach is illustrated in a challenging natural image segmentation task, showing good performance with respect to the literature.

EO688 Room R24 ADVANCES IN MULTIVARIATE BAYESIAN METHODS Chair: Ioanna Manolopoulou

E0356: **Bayesian causal structural learning with zero-inflated Poisson Bayesian networks**
Presenter: **Yang Ni**, Texas A&M University, United States

Multivariate zero-inflated count data arise in a wide range of areas such as economics, social sciences, and biology. To infer causal relationships in zero-inflated count data, we propose a new zero-inflated Poisson Bayesian network (ZIPBN) model. We show that the proposed ZIPBN is identifiable with cross-sectional data. The proof is based on the well-known characterization of Markov equivalence class which applies to other distribution families. For causal structural learning, we introduce a fully Bayesian inference approach which exploits the parallel tempering Markov chain Monte Carlo algorithm to explore the multi-modal network space efficiently. We demonstrate the utility of the proposed ZIPBN in causal discoveries for zero-inflated count data by simulation studies with comparison to alternative Bayesian network methods. Additionally, real single-cell RNA-sequencing data with known causal relationships will be used to assess the capability of ZIPBN for discovering causal relationships in real-world problems.

E0372: **Estimating heterogeneous causal effects in time series settings with staggered adoption**
Presenter: **Joseph Antonelli**, University of Florida, United States
Co-authors: Brenden Beck

Communities often self select into implementing a regulatory policy and adopt the policy at different time points. Researchers are interested in (1) evaluating the impact of the policy, and (2) understanding what types of communities are most impacted by the policy, raising questions of heterogeneous treatment effects. We develop novel statistical approaches to study the causal effect of policies implemented at the community level. Using techniques from high-dimensional Bayesian time-series modeling, we estimate treatment effects by predicting counterfactual values of what would have happened in the absence of the policy. We couple the posterior predictive distribution of the treatment effect with flexible modeling to identify how the impact of the policy varies across time and community characteristics. This allows us to identify effect modifying variables and capture nonlinear heterogeneous treatment effects. Importantly, our approach is robust to unmeasured confounding bias. Our methodology is motivated by studying the effect of neighborhood policing on arrest rates in New York City. Using realistic simulations based on the policing data in New York City, we show our approach produces unbiased estimates of treatment effects with valid measures of uncertainty. Lastly, we find that neighborhood policing decreases arrest rates after treatment adoption, but has little to no effect on other outcomes of interest.

E0858: **A comprehensive Bayesian framework for envelope models**
Presenter: **Saptarshi Chakraborty**, State University of New York at Buffalo, United States
Co-authors: Zhihua Su

The envelope model aims to increase efficiency in multivariate analysis. It has been used in many contexts, including linear regression, generalized linear models, matrix/tensor variate regression, reduced rank regression, and quantile regression. It has shown the potential to provide substantial efficiency gains. Virtually all of these advances, however, have been made from a frequentist perspective, and the literature addressing envelope models from a Bayesian point of view is sparse. The objective is to propose a comprehensive Bayesian framework that is applicable across various envelope model contexts. The proposed framework aids the straightforward interpretation of model parameters and allows easy incorporation of prior information. We provide a simple block Metropolis-within-Gibbs MCMC sampler for the practical implementation of our method. Simulations and data examples show impressive efficiency gains over standard Bayesian regression methods.

E0902: **Quantile regression under spatial noise correlation**
Presenter: **Surya Tokdar**, Duke University, United States

Quantile regression is widely adopted for regression analyses in ecology, economics, education, public health and climatology. In QR, one replaces the standard regression equation of the mean with a similar equation for a quantile at a given quantile level of interest. But the real strength of QR lies in the possibility of analyzing any quantile level of interest, and perhaps more importantly, contrasting many such analyses against each other with fascinating consequences. Despite the popularity of QR, it is only recently that an analysis framework has been developed which transforms the four-decade-old idea into a model-based inference and prediction technique in its full generality. In doing so, the new joint estimation framework has opened doors to many important advancements of the QR analysis technique to address additional data complications. The focus

is on extending quantile regression to analyzing geolocated data while adjusting for spatial noise correlation. We will show how the considered modeling framework allows a new interpretation of 'noise' in the QR context, and how one may seamlessly generalize this estimation framework to account for a wide variety of noise dependency, including tail dependence appropriate for heavy-tailed data. The new method is applied to air quality and wildfire risk analyses.

EC798 Room R22 CONTRIBUTIONS IN STOCHASTIC PROCESSES Chair: Arnaud Gloter

E0783: **An unified stochastic hyperbolastic model**
Presenter: **Antonio Barrera**, Universidad de Malaga, Spain
Co-authors: Patricia Roman-Roman, Francisco Torres-Ruiz

The hyperbolastic growth curves have been successfully applied to describe many dynamical phenomena in biosciences, in particular, deterministic growth of cell populations. There are three different types, two of them being extensions of classic growth models such as the logistic and the Weibull curves. Stochastic counterparts of these models have been built in order to describe random influence in growth phenomena. Therefore, some problems involving parameter estimation have been addressed by applying both analytic and approximated methods. The main goal is to establish a common framework for all hyperbolastic models by considering similar strategies to address issues related to inference. In particular, bounded parametric spaces and initial solutions of likelihood equations are discussed. On the other hand, the relation between hyperbolastic models and classic growth curves may lead to some questions about the theory of generalized growth models.

E1139: **Modelling self-excitation of extreme returns in financial markets: An AR-GARCH with Hawkes Jumps approach**
Presenter: **Steve Yang**, Stevens Institute of Technology, United States
Co-authors: Anqi Liu

Extreme value theory (EVT) and Hawkes processes in the AR-GARCH framework are applied to model the tail risk clustering effect. The proposed model improves forecasts of the timing of extreme returns and is particularly useful for downside risk analysis. Due to a large parameter set, we propose a two-step calibration method to estimate the model. We apply this model on 90 stocks, including both large-caps and small-caps, in nine industry sectors. The in-sample experiments show a strong self-exciting of negative extremal of AR-GARCH residuals and it is well captured using our model. The value-at-risk forecasting experiments over the past 25 years confirm that the proposed model produces accurate downside risk estimations. More importantly, the proposed model provides more stable risk analysis results during the market crisis than other existing benchmark models.

E0793: **Detecting periodicity from the trajectory of a random walk in random environment**
Presenter: **Jean Vaillancourt**, HEC Montreal, Canada
Co-authors: Bruno N Remillard

For nearest neighbour univariate random walks in a periodic environment, where the probability of moving depends on a periodic function, we show how to estimate the period and the function. For random walks in non-periodic environments, we find that the asymptotic limit of the estimator is constant in the ballistic case when the random walk is transient, and the law of large numbers holds with a non zero limit. Numerical examples are given in the recurrent case, as well as in the sub-ballistic case (where the random walk is transient, but the law of large numbers yields a zero limit).

E0530: **Dependence under random time-varying network distances with an application to Cox-Processes**
Presenter: **Alexander Kreiss**, KU Leuven, Belgium

Multivariate stochastic processes indexed either by vertices or pairs of vertices of a dynamic network are considered. Under a dynamic network we understand a network with a fixed vertex set and an edge set which changes randomly over time. The aim is to conduct inference in models for this type of data. For real-world applications, it is important to allow that processes of adjacent pairs (or adjacent vertices) may be dependent. Since networks are often changing over time, the notion of adjacency is dynamic and as a consequence also the dependence should be dynamic. We will thus assume that the spatial dependence structure of the processes conditional on the network behaves in the following way: Close vertices (or pairs of vertices) are dependent, while we assume that the dependence decreases conditionally on that the distance in the network increases. We make this intuition mathematically precise by considering three concepts based on correlation, beta-mixing with time-varying beta-coefficients and conditional independence. These concepts allow proving weak-dependence results, e.g. an exponential inequality, which might be of independent interest. In order to demonstrate the use of these concepts in an application we study the asymptotics (for growing networks) of a goodness of fit test in a dynamic interaction network model based on a Cox-type model for counting processes. This model is then applied to bike-sharing data.

EC801 Room R25 CONTRIBUTIONS IN NONPARAMETRIC STATISTICS AND RESAMPLING Chair: Anneleen Verhasselt

E1048: **On the mysteries of resampling for matching estimators**
Presenter: **Christopher Walsh**, TU Dortmund University, Germany
Co-authors: Carsten Jentsch, Shaikh Tanvir Hossain

In a highly influential paper showed that, in general, the conditional variance of an Efron-type bootstrap for the matching estimator does not converge to the correct limit. However, they also conjecture that the asymptotic variance should be consistently estimable by using a wild bootstrap or an M-out-of-N bootstrap. We prove that the conditional variance of: (i) a wild-type bootstrap procedure recently proposed in general does not converge to the correct limit – either in the setting considered for the ATET or in a slightly modified design for the ATE; (ii) an M-out-of-N-type bootstrap estimator does converge to the correct limit in expectation in the setting considered previously. Intuitively, the Efron-type bootstrap estimator fails, because it does not replicate the matching algorithm correctly due to the presence of ties in the resamples. This is not the case for the proposed M-out-of-N-type bootstrap as it does not contain any observations more than once with probability one. Extensive simulations support our theoretical findings for the wild-type bootstrap and the M-out-of-N-type bootstrap.

E0978: **The effect of smoothing on tests based on regression residuals**
Presenter: **Natalia Perez-Veiga**, Universidade de Vigo, Spain
Co-authors: Juan-Carlos Pardo-Fernandez

In the regression context, several testing procedures have been designed based on the distribution of the residuals. Typically, the distribution functions involved in test statistics are estimated using the empirical distribution function, which, because of its discontinuous nature, can result in poor estimates of the true distribution function. Through a simulation study, it is shown that replacing the empirical distribution function by its smoothed version results in a power improvement. This methodology is applied to the particular cases of goodness-of-fit tests and comparison of regression curves.

E1049: **Nonparametric moment-based estimation of simulated models without optimization**
Presenter: **Raffaello Seri**, University of Insubria, Italy
Co-authors: Mario Martinoli

A new method for the estimation of simulated models is presented. It exploits a nonparametric sieve regression estimated through OLS to find the parameters of a simulation model producing statistics that are close to the ones obtained in real-world data. The simulation model is run for several values of the parameters, statistics are computed on each run, and the function linking the generated statistics and the associated parameters is estimated nonparametrically. Estimates of the parameters are then obtained through the previous nonparametric estimate using the

real-world statistics as explanatory variables. At odds with simulated minimum-distance techniques (e.g., indirect inference and simulated method of moments), our framework does not involve any objective function, and no optimization algorithm is required. The full asymptotic theory of the estimator is explicitly and rigorously characterized, including the order of the bias, confidence intervals and hypotheses tests. The approach is evaluated through a small simulation study and the estimation of an agent-based computational model in which the evolutionary dynamics of the financial market are driven by agents with heterogeneous beliefs.

E0296: **Nonparametric C- and D-vine based quantile regression**
Presenter: **Marija Tepegjozova**, Technical University Munich, Germany
Co-authors: Claudia Czado, Gerda Claeskens, Jing Zhou

Quantile regression is a field with steadily growing importance in statistical modeling. It is a complementary method to linear regression, since computing a range of conditional quantile functions provides more accurate modelling of the stochastic relationship among variables, especially in the tails. We introduce a novel non-restrictive and highly flexible nonparametric quantile regression approach based on C- and D-vine copulas. Vine copulas allow for separate modeling of marginal distributions and the dependence structure in the data. They can be expressed through a graph theoretical model given by a sequence of trees. This way, we obtain a quantile regression model, that overcomes typical issues of quantile regression such as quantile crossings or collinearity, the need for transformations and interactions of variables. Our approach incorporates a two-step ahead ordering of variables, by maximizing the conditional log-likelihood of the tree sequence, while taking into account the next two tree levels. Further, we show that the nonparametric conditional quantile estimator is consistent. The performance of the proposed methods is evaluated in both low- and high-dimensional settings using simulated and real-world data. The results support the superior prediction ability of the proposed models.

CO427 Room R03 ASSET PRICING WITH NON-STANDARD RISKS Chair: Mirco Rubin

C0298: **Smart stochastic discount factors**
Presenter: **Alberto Quaini**, University of Geneva, Switzerland
Co-authors: Fabio Trojani, Sofonias Alemu Korsaye

A unifying theoretical framework is developed for selecting model-free Stochastic Discount Factors (SDFs) in arbitrage-free markets under general convex constraints on pricing errors, and show that such SDFs arise in a wide range of economies featuring, e.g., various forms of frictions, ambiguous asset payoffs, asymptotic no-arbitrage conditions under Ross' Arbitrage Pricing Theory (APT), or a need for regularization in large asset markets. We introduce a new family of minimum variance SDFs incorporating APT pricing error bounds, which are designed to optimize the tradeoff between pricing accuracy and the SDF ability to comove with systematic asset return risks. Empirically, we find that a model-free adaptation of an SDF under the CAPM, which exactly prices market risk but otherwise constrains the amount of mispricing across assets with a model-free APT pricing error bound, generates an optimal tradeoff.

C0480: **A structural model of market friction with time-varying volatility**
Presenter: **Giuseppe Buccheri**, University of Rome Tor Vergata, Italy
Co-authors: Stefano Grassi, Giorgio Vocalelli

The problem of extracting the volatility of a financial security is considered when its prices are not frequently updated over time. A model of price formation is proposed in which the observed price varies only if the value of the information signal is large enough to guarantee a profit in excess of transaction costs. Using transaction data only, we extract: (i) the conditional volatility of the underlying security, which is thus cleaned out by market frictions, (ii) an estimate of transaction costs. We apply the model to a large dataset of intraday data. The analysis reveals that, when correcting for transaction costs, the risk of illiquid securities is substantially different from what predicted by traditional volatility models.

C0618: **The effect of personal taxes on investment decisions and stock returns**
Presenter: **Alex Kontoghiorghes**, Queen Mary University of London, United Kingdom

The aim is to study how the tax rate faced by retail investors for holding publicly listed companies affects the stock demand, stock return, and dividend policy of those companies. A change in legislation in 2013 is exploited which allowed stocks listed on the Alternative Investment Market (AIM), a sub-market of the London Stock Exchange, to be held in capital gains and dividend tax-exempt investment accounts for the first time. After the tax cut, stock demand increased, stock returns decreased, and dividends increased. AIM stocks can also be uniquely exempt from a 40% inheritance tax. Hence to rationalize results, a life-cycle model is introduced which accounts for different taxes and a bequest motive. The results demonstrate the importance of personal taxes for both investors and companies.

C0612: **Equity tail risk in the treasury bond market**
Presenter: **Mirco Rubin**, EDHEC - Nice, France
Co-authors: Dario Ruzzi

The effects of equity tail risk on the US government bond market are quantified. We estimate equity tail risk with option-implied stock market volatility that stems from large negative price jumps and we assess its value in (i) reduced-form predictive regressions for Treasury returns, and (ii) an affine term structure model. We find that the left tail volatility of the stock market significantly predicts one-month excess returns on Treasuries both in- and out-of-sample. The incremental value of employing equity tail risk as a return forecasting factor can be of economic importance for a mean-variance investor trading bonds. The estimated term structure model shows that equity tail risk is priced in the US government bond market. Consistently with the theory of flight-to-safety, we find that when the perception of tail risk is higher (i) Treasury prices increase, and (ii) funds are flowing out of the equity market and (iii) into the bond market. Our results concerning the predictive power and pricing of equity tail risk extend to major government bond markets in Europe.

CO071 Room R04 CONTRIBUTIONS IN BAYESIAN ECONOMETRICS Chair: Sylvia Fruehwirth-Schnatter

C0751: **Estimation for univariate and bivariate reinforced urn processes under left-truncation and right-censoring**
Presenter: **Luis Antonio Souto Arias**, Centrum Wiskunde and Informatica-Delft University of Technology, Netherlands
Co-authors: Pasquale Cirillo, Cornelis W Oosterlee

Reinforced Urn Processes (RUPs) represent a flexible class of Bayesian nonparametric models suitable for dealing with possibly right-censored and left-truncated observations. A reliable estimation of their hyper-parameters is, however, missing in the literature. We, therefore, propose an extension of the Expectation-Maximization (EM) algorithm for RUPs, both in the univariate and the bivariate case. Furthermore, a new methodology combining EM and the prior elicitation mechanism of RUPs is developed: the Expectation-Reinforcement algorithm. Using a well-known Canadian dataset, the performances of both algorithms are studied in the context of joint and last survivor annuity pricing.

C1042: **Bayesian multilevel hidden Markov models: A reliability guideline for applied researchers**
Presenter: **Sebastian Mildiner Moraga**, Utrecht University, Netherlands
Co-authors: Emmeke Aarts

The popularity of hidden Markov models (HMMs) in the econometric, social and behavioural literature has been steadily increasing lately. Moving beyond standard statistical tests, HMMs offer a statistical environment to optimally exploit the information present in real-time behavioural data, for example, uncovering the dynamics of behaviour. For this type of data, multilevel hidden Markov models (MHMMs) are a particularly good fit: they allow for the accommodation of variability between subjects and the estimation of both subject-level and group-level parameters. To create

guidelines for applied researchers, we assessed the effect of three researcher controlled factors -the number of subjects, the number of occasions and the number of dependent variables used to train the model- and two data quality conditions -noisiness and overlapping of the conditional distributions- on the estimation performance of a multinomial Bayesian MHMM. Our results reveal that increments in the number of subjects and the number of dependent variables included are more beneficial for the estimation performance than increments on the number of occasions. These effects are consistent across the levels of complexity in the conditional distributions. We conclude that measuring multivariate rather than univariate data results in the most cost-effective gains in estimation performance and the likelihood of convergence.

C0286: **The role of information demand in oil and gas markets**
Presenter: **Georgios Papapanagiotou**, University of Macedonia, Greece
Co-authors: Georgios Bampinas, Theodore Panagiotidis

Two indices are presented that reflect the information demand for oil and shale gas. They are employed to analyse the behaviour of the energy commodities market. To construct the two indices we use the Search Volume Index data from Google search data for a set of search terms related to oil and shale gas. A two-step dynamic factor model is used that is based on Kalman filtering which captures the main bulk of co-movement between the data. Next, we employ a Bayesian time varying parameter VAR (TVP-VAR) to study the response of the energy commodities market to shocks in information demand and vice versa.

C0352: **Bayesian inference in censored linear regression model with AR(p) errors**
Presenter: **Rodney Sousa**, Universidade de Aveiro, Portugal

Censored linear regression models are a class of models which allow the dependent variable to be censored. The problem of estimating a censored linear regression model with autocorrelated errors, CLR-AR, may arise in many environmental and social studies. The likelihood function for this model is very complex, hindering the use of Maximum Likelihood methods in real problems, especially when the number of censored observations in the sample is large. We consider a Bayesian approach to estimate the CLR-AR. We propose a Gibbs sampler with Data Augmentation algorithm in which each augmented datum is, in fact, the mean of multiple simulated values, GDA-MMS. The performance of the algorithm is analyzed in a simulation study. The results indicate that the estimates show good accuracy and Bayesian consistency, even in scenarios where the proportion of censored observations is large. The approach is also illustrated in an empirical dataset regarding cloud ceiling height, where 41% of the observations are censored.

CO061 Room R06 APPLIED NETWORK ANALYSIS IN EMPIRICAL FINANCE Chair: Massimo Guidolin

C0330: **Stock market as a network**
Presenter: **Marcello Esposito**, Universita Cattaneo, Italy

Among the statistical techniques used to describe the behaviour of the financial markets, one of the most promising is based on the network analysis of the stock market. In this framework, the stock market is represented as a graph with nodes (the single stocks), edges (connections between stocks), and attributes (industry classification, volumes). The application of network analysis to the stock market is not new, but in the literature the market graph has been mainly derived from the correlation matrix of the stock prices. This is a limitation and the risks are to express in different words what traditional financial econometrics has already said about the returns distribution. For this reason, we integrated the analysis and built the market graph with new type of data taken from the observation of the information gathering activity performed by retail investors through the Googles search engine. We focussed the attention to financial crises, when a shock hits the economy in such a profound way that almost all the parameters entering the pricing equation of stocks must be reassessed. Those periods are relatively rare and short. They are characterised by extremely high levels of volatility and correlation. In these moments, searching for new information becomes of paramount importance. And then it is in these moments that we expect to observe more neatly the working of the underlying network.

C0325: **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 previously, which is based on realized, backwards-looking volatilities instead. While both measures show evidence of in-sample predictive power, only the option-implied measure can produce out-of-sample forecasts that outperform a simple historical mean benchmark.

C0343: **Cyberattacks and cross-market bitcoin prices: Network analysis**
Presenter: **Ahmad Maaitah**, University of Southampton, United Kingdom
Co-authors: Mauro Costantini, Tapas Mishra

The volatilities across Bitcoin markets interconnected, especially during episodes of cyberattacks are investigated. A significant premise concerns major cyberattacks, when the highly interconnected volatile Bitcoin system is exposed to systemic risk and unfounded market exploitations, leading further to a depth in volatility across this market as well as other interdependent markets. Dynamic conditional volatility measures across six major Bitcoin markets are estimated. Network analysis is performed to characterise the strength and direction of interconnectedness. We find evidence of a significant movement of volatilities across the markets over time. Such volatilities appear to depict a trend and high interconnectedness when the number and magnitude of cyberattacks rise. This implies a major inefficiency of the Bitcoin market system that can be exploited to gain a major advantage of arbitrage.

C0326: **Leveraged loans, systemic risk and network interconnectedness**
Presenter: **Ana Sina**, University of Reading, United Kingdom
Co-authors: Monica Billio, Alfonso Dufour, Simone Varotto

Especially among those who lived the global financial crisis of 2007-2008, it is likely that few regulators will instinctively be careless about the above pre-crisis level achieved by the global syndicated leveraged loans. We introduce global syndicated leveraged loans to study the relationship between systemic risk, syndicated leveraged loans, and network interconnectedness. We analyse the U.S. and European syndicated loans that together account for almost 80% of the global syndicated database during the period from 1988 to 2019. By distinguishing between leveraged and other syndicated loans, we develop a novel measure of systemic risk as the ratio between leveraged and total loans that each lead arranger holds. We make the following contributions. First, we show that leveraged loans have already exceeded the pre-financial crisis level, which may pose financial stability concerns. Second, we relate our novel measure with the systemic risk of each financial institution and find a significant correlation, which may suggest that this new measure is employable as an early-warning. Finally, since the syndicated loan market can be represented as a network where the linkages and nodes are based on real collaborations among lead arrangers, we show that the financial institutions that have a larger proportion of leveraged loans in their portfolio are the most central in the network and this may cause possible implications for the future financial stability.

CO660 Room R08 THE MACROECONOMIC PROPAGATION OF SHOCKS Chair: Filippo Ferroni

C0464: **The covid-induced uncertainty shocks**
Presenter: **Mirela Sorina Miescu**, Lancaster University, United Kingdom
Co-authors: Raffaele Rossi
The causal effects of Covid-induced economic uncertainty within a daily structural VAR of the US, over the sample January-July 2020, are estimated. The key identifying assumption is that uncertainty shocks are heteroskedastic, so that they have especially high variance on days when there are important announcements about the pandemic. We find that Covid-induced uncertainty shocks lead to a significant contraction of economic and financial indicators. Monetary policy reacts promptly to this increase in uncertainty. We also find important distributional effects: in the face of an increase of Covid-induced uncertainty, low-income households suffer a contraction in employment twice as large as the high-income households, while expenditure of the richest top quartile contracts 40% more than that of the bottom quartile. Finally, we show that industries which rely on face-to-face interactions, such as entertainment and hospitality, see a reduction in their revenues more than three times larger than industries that can operate remotely, such as business services.

C0590: **Estimating hysteresis effect**
Presenter: **Francesco Furlanetto**, Norges Bank, Norway
The standard Blanchard-Quah decomposition is extended to enable fluctuations in aggregate demand to have a long-run impact on the productive capacity of the economy through hysteresis effects. These demand shocks are found to be quantitatively important in the US, in particular if the Great Recession is included in the sample. Demand-driven recessions lead to a permanent decline in employment while output per worker is largely unaffected. The negative impact of a permanent decline in investment (including R&D investment) on productivity is compensated by the fact that the least productive workers are disproportionately hit by the shock and exit the labor force.

C0588: **Risk-sharing channels in OECD countries: A heterogeneous panel var approach**
Presenter: **Pilar Poncela**, Universidad Autonoma de Madrid, Spain
Co-authors: Pierfederico Asdrubali, Soyoung Kim, Filippo Pericoli
The aim is to improve upon the existing empirical literature on international risk sharing under three dimensions. First, we generalize dynamic multi-equation approaches to the estimation of risk-sharing channels, by adopting a Heterogeneous Panel VAR model. Within this framework, the coefficients representing the extent of risk sharing achieved through the different mechanisms are allowed to vary across countries. Second, we introduce two new risk-sharing channels, namely, government consumption and the real exchange rate, which allow us to investigate the role of fiscal policy and international price adjustments in the absorption of macroeconomic shocks. Third, we establish a better link between the channels empirical model and a theoretical formulation of the risk-sharing condition, which allows for PPP violations. The empirical analysis, for a set of 21 OECD countries over the 1960-2016 period, contributes to identifying the geographical structure and dynamics of risk-sharing channels and to describing their evolution in the latest half-century. For the OECD sample as a whole, we confirm through 2016 the strong smoothing role played by credit markets and the small degree of risk-sharing achieved through factor incomes. Another noteworthy result is the negative risk-sharing effect of the real exchange rate, driven by the dis-smoothing role played by the movements of the nominal exchange rate, only partially offset by relative price adjustments.

C1089: **Tracking economic growth during the Covid-19: A weekly indicator for Italy**
Presenter: **Simone Emiliozzi**, Banca d'Italia, Italy
Co-authors: Davide Delle Monache, Andrea Nobili
Following the breakout of the COVID-19 pandemic, economic forecasting has become more complex. One way to address these challenges is to exploit the information content of high-frequency variables to construct a synthetic and timely indicator of the business cycle. A weekly economic activity indicator for the US economy has been proposed. We recently developed a preliminary version of an Italian Weekly Economic Index (ITWEI), which was particularly useful for policy analysis in the current circumstances. In its present version, the ITWEI is obtained as the principal component of ten series including genuinely weekly variables as well as monthly indicators to be disaggregated at a weekly frequency. The aim is twofold: (i) to enlarge the set of weekly indicators to exploit more high-frequency information such other payment system data, some financial variables, electronic receipts, sentiment and uncertainty indicators stemming from textual data; (ii) to refine the methodology using a unified state space approach to deal with mixed frequency and missing observations.

CO289 Room R19 ADVANCES IN ROBUST ESTIMATION AND INFERENCE: THEORY AND APPLICATIONS II Chair: Rustam Ibragimov

C1038: **Diversity in news recommendations using contextual bandits**
Presenter: **Alexander Semenov**, University of Florida and Saint Petersburg State University, United States
Co-authors: Gaurav Pandey, Maciej Rysz, Guanglin Xu
Contextual bandit techniques have recently been used for generating personalized user recommendations in situations where collaborative filtering based algorithms may be inefficient. They are often used in cases when input data are dynamically changing as new users, and content items constantly change. One such setting involves recommending news articles to users based on context, i.e., user and article features. Contextual bandit methods sequentially select articles for recommendation to a user and continuously modify their strategies to present users with articles that maximize clicks. However, exclusively focusing on maximizing the number of clicks can lead to over-exposure of certain articles, while under-representing others. In an era of the ever-growing demand for digital news delivery, this, in turn, invokes the important notion of presenting news content to users in a "socially responsible" way. We present a technique based on the contextual bandit framework that, in addition to maximization of the click rate, also considers the historical frequency of an article as the "cost" associated with recommending it. It is demonstrated that this approach results in a more balanced distribution and a diverse set of recommended articles. Experiments utilizing a benchmark news dataset demonstrate the trade-off between clicks and diversity of recommended articles.

C1097: **Robust inference on income inequality: t-statistic based approaches**
Presenter: **Rustam Ibragimov**, Imperial College London and St. Petersburg State University, United Kingdom
Co-authors: Paul Kattuman, Anton Skrobotov
Empirical analyses on income and wealth inequality often face the difficulty that the data are heterogeneous, heavy-tailed or correlated in some unknown fashion. The focus is on applications of the recently developed robust t-statistic methods in the analysis of inequality measures and their comparisons under the above problems. In particular, a robust large sample test on equality of two parameters of interest (e.g., a test of equality of inequality measures in two regions or countries) is conducted as follows: The data in the two samples dealt with is partitioned into fixed numbers q_1, $q_2 > 1$ (e.g., $q1 = q2 = 2, 4, 8$) of groups, the parameters (inequality measures) are estimated for each group, and inference is based on a standard two-sample t-test with the resulting q_1, q_2 group estimators. This results in valid inference under general conditions that group estimators of parameters (e.g., inequality measures) considered weakly converge, at an arbitrary rate, to independent mixed normal random variates. The conditions are typically satisfied in empirical applications even under pronounced heavy-tailedness and heterogeneity and possible dependence in observations. The methods complement and compare favorably with other inference approaches available in the literature. The use of robust inference approaches is illustrated by an empirical analysis of income inequality measures and their comparisons across different regions in Russia.

C0986: **On the consistency of nonparametric bootstrap for inference on high-quantile, tail index, and tail probability**
Presenter: **Svetlana Litvinova**, Monash University and St. Petersburg State University, Australia
Co-authors: Mervyn Silvapulle

The full-sample bootstrap is shown to be asymptotically valid for constructing confidence intervals for high-quantiles, tail probabilities, and other tail parameters of a univariate distribution. This resolves the doubts that have been raised about the validity of such bootstrap methods. In our extensive simulation study, the overall performance of the bootstrap method was better than that of the standard asymptotic method. The method of proof is likely to be useful for studying nonparametric bootstrap for inference about extreme tail events more generally.

C1101: **Predictability of cryptocurrency returns: Evidence from robust tests**
Presenter: **Siyun He**, University of Michigan, United States
Co-authors: Rustam Ibragimov

The purpose is to provide a comparative empirical study of predictability of cryptocurrency returns and prices using econometrically justified robust inference methods. We present a robust econometric analysis of predictive regressions incorporating factors including cryptocurrency momentum, stock market factors, acceptance of Bitcoin, and Google trends measure of investors' attention. Due to inherent heterogeneity and dependence properties of returns and other time series in financial and crypto markets, we provide the analysis of the predictive regressions using HAC standard errors. We further present the analysis of the predictive regressions using recently developed t-statistic robust inference approaches. We provide comparisons of robust predictive regression estimates between different cryptocurrencies and their corresponding risk and factor exposures. In general, the number of significant factors decreases as we use more robust t-tests, and the t-statistic robust inference approaches appear to perform better than the t-tests based on HAC standard errors in terms of pointing out interpretable economic conclusions. The results emphasise the importance of the use of robust inference approaches in the analysis of economic and financial data affected by the problems of heterogeneity and dependence.

CG336 Room R02 CONTRIBUTIONS IN TIME SERIES AND FORECASTING Chair: Christian Conrad

C0188: **On the estimation of value-at-risk and expected shortfall at extreme levels**
Presenter: **Jingqi Pan**, University of Reading, United Kingdom
Co-authors: Emese Lazar, Shixuan Wang

Two dynamic semi-parametric models that estimate Value-at-Risk (VaR) and Expected Shortfall (ES) are generalized, specifically the one-factor GAS model and the Hybrid GAS/GARCH model, to enhance them for risk estimation at extreme levels (corresponding to very low values of alpha). We achieve this by simultaneously estimating VaR and ES for multiple levels of alpha. We found that this approach improves on the risk estimation for low alpha values by having a unique hidden process that drives risk estimates for multiple levels of alpha. The simulation results indicate that both generalized models are better than their corresponding benchmarks in terms of estimated loss, forecast loss and the percentage backtest rejections for extreme values of alpha. We demonstrate the applicability of the generalized models on daily returns on four international equity indices, and the empirical results show the superior performance of the generalized models.

C0502: **Assumptions and macroeconomic forecasts: Disagreement, revisions and forecast errors**
Presenter: **Katja Heinisch**, Halle Institute for Economic Research, Germany
Co-authors: Alexander Glas

Using data from the European Central Bank's Survey of Professional Forecasters, the role of ex-ante conditioning assumptions for macroeconomic forecasts is analyzed. In particular, we test to which extent the heterogeneity, updating and ex-post performance of predictions for inflation, real GDP growth and the unemployment rate can be related to assumptions about oil prices, exchange rates, interest rates and wage growth. Our findings indicate that experts predict macroeconomic outcomes in line with well-known theoretical relationships such as the Phillips curve, Okun's Law or the Taylor rule. Inflation forecasts are closely associated with oil price assumptions, whereas interest rate assumptions are used primarily to forecast output growth and unemployment. Exchange rate and wage growth assumptions also play a role in shaping forecasts, albeit less so than oil prices and interest rates. The findings indicate that survey participants can improve the accuracy of their macroeconomic predictions by reducing assumption errors. These results contribute to a better understanding of the expectation formation process of economic agents.

C1051: **Entropic tilting for macroeconomic variables: The role of asymmetrically distributed survey forecasts**
Presenter: **Anastasia Allayioti**, University of Warwick, United Kingdom

There is a growing interest in incorporating external information extracted from a survey of professional forecasters into real-time macroeconomic predictions from vector autoregressive (VAR) specifications. The method of entropic tilting achieves this by modifying the baseline VAR distribution such that it matches certain moment conditions. Existing papers adopting this methodology focus on the first two moments of the forecast distribution (mean and variance). By implicitly assuming that the first two moments summarize all required information, these papers restrict their attention to a symmetric environment. We propose a modification to the standard relative entropy approach which allows for asymmetry in the macroeconomic variables and explores the predictive content of higher-order moments. The proposed methodology involves tilting the VAR distribution towards an aggregate survey forecast density that has been appropriately reshaped to match the non-Gaussian features of the sample data. We illustrate this methodology with an application examining real-time forecasts for four U.S. macroeconomic variables. We consider a variety of VAR models, ranging from time-varying volatility to non-Gaussian errors. Results across models indicate meaningful gains in terms of both point and density forecast accuracy relative to individual multivariate specifications and existing forecasting methods that blend model-based forecasts with external judgement.

C0690: **Retail investors' trading activity and the predictability of stock return correlations**
Presenter: **Daniele Ballinari**, University of St Gallen, Switzerland

Considerable theoretical and empirical evidence links price comovements with the behavior of retail investors. Nevertheless, when predicting stock return correlations, research has focused on the leverage effect. We propose a new model of realized covariances that allows exogenous predictors to influence the correlation dynamics while ensuring the predicted matrices' positive definiteness. Using this model, the predictive power of retail investors' sentiment and attention for the correlations of 35 Dow Jones stocks is analyzed. We find retail investors' attention to have predictive power for return correlations, especially for longer forecasting horizons and during the COVID-19 pandemic. Value-at-risk forecasts confirm these results.

CG116 Room R07 CONTRIBUTIONS IN STATISTICS AND ECONOMETRICS FOR THE COVID-19 PANDEMIC Chair: Doris Behrens

E1181: **To freeze or not to freeze: Epidemic prevention and control in the DSGE model with agent-based epidemic component**
Presenter: **Jagoda Kaszowska-Mojsa**, Polish Academy of Sciences/Cracow University of Economics, Poland
Co-authors: Przemyslaw Wlodarczyk, Przemyslaw Wlodarczyk

The ongoing epidemic of COVID-19 raises numerous questions concerning the shape and range of state interventions, that are aimed at reduction of the number of infections and deaths. The lockdowns, which became the most popular response worldwide, are assessed as being an outdated and economically inefficient way to fight the disease. However, in the absence of efficient cures and vaccines, they lack viable alternatives. We assess the economic consequences of epidemic prevention and control schemes that were introduced to respond to the COVID-19 outburst. The analyses

report the results of epidemic simulations obtained with the agent-based modeling methods under different response schemes and use them in order to provide conditional forecasts of standard economic variables. The forecasts are obtained from the DSGE model with labour market component.

E1176: **Benford law and the reliability of COVID-19 data in western Balkan countries**

Presenter: **Eralda Gjika**, University of Tirana, Albania

Co-authors: Lule Basha

Almost 1 year after the outbreak of the covid-19 pandemic the figures published by the institutions responsible for information in some states leave doubt in their statements. There are some probability tests used for the reliability of the information, among which Benford's law is the one also used in fraud detection. The focus is on the Western Balkan countries where the appearance of COVID-19 was delayed by almost 2 months. We have analyzed through Benford law the reliability of the figures published each day by official state institutions: new cases displayed and deaths by COVID-19. The western Balkan region is considered, and for a better view of the results, we have also studied the situation in some of the most vulnerable countries during October 2020. What is noticed within the analysis of these data is the evidence of manipulation in published data in western Balkan countries and also in some other states which may be affected by many factors such as the number of tests per day. Policymakers should take into consideration veracity of the information before undertaking policies which can have consequences for the country and its population.

C0437: **Pandemic shocks and household spending**

Presenter: **David Finck**, University of Giessen, Germany

The response of daily household spending to the unexpected component of the COVID-19 pandemic, which we label as pandemic shock, is studied. Based on daily forecasts of the number of fatalities, we construct the surprise component as the difference between the actual and the expected number of deaths. We allow for state-dependent effects of the shock depending on the position on the curve of infections. Spending falls after the shock and is particularly sensitive to the shock when the number of new infections is strongly increasing. If the number of infections grows moderately, the drop in spending is smaller. We also estimate the effect of the shock across income quartiles. In each state, low-income households exhibit a significantly larger drop in consumption than high-income households. Thus, consumption inequality increase after a pandemic shock. Our results hold for the US economy and the key US states. The findings remain unchanged if we choose alternative state-variables to separate regimes.

C0342: **COVID-19 impact on stock prices: The sector analysis**

Presenter: **Patrycja Chodnicka - Jaworska**, University of Warsaw, Poland

Co-authors: Piotr Jaworski

The aim is to examine the impact of information on lockdown connected with COVID-19 on stock prices according to the sector and country divisions. The hypotheses are: it is observed a strong negative impact of the lockdown announcement on stock prices at the time of its announcement. The reaction of the stock prices on the lockdown announcement is varied in particular countries and sectors. Panel event models were used to verify the hypothesis. The study used data from the Thomson Reuters database for the period from 02.01.2019 - 10.05.2020. The analysis was based on the papers and reports on COVID-19 and the literature on behavioural finance.

Sunday 20.12.2020 15:05 - 16:20 Parallel Session J – CFE-CMStatistics

EO429 Room R11 NONLINEAR METHODS IN FUNCTIONAL DATA ANALYSIS **Chair: Alexander Petersen**

E0222: **Intrinsic Riemannian functional data analysis**
Presenter: **Zhenhua Lin**, National University of Singapore, Singapore
Co-authors: Fang Yao

Data of random paths on a Riemannian manifold is often encountered in real-world applications. Examples include trajectories of bird migration, the dynamics of brain functional connectivity, etc. To analyze such data, a framework of intrinsic Riemannian functional data analysis is developed, which provides a rigorous theoretical foundation for statistical analysis of random paths on a Riemannian manifold. The cornerstone of the framework is the Hilbert space of vector fields along a curve on the manifold, based on which principal component analysis and Karhunen-Love expansion for Riemannian random paths are then established. The framework also features a method for proper comparison of vector fields along different curves, which paves the way for intrinsic asymptotic analysis of estimation procedures for Riemannian functional data analysis. Built on intrinsic geometric concepts such as vector field, Levi-Civita connection and parallel transport on Riemannian manifolds, the proposed framework is able to properly handle intrinsic geometric concepts. Based on the framework, functional linear regression models for Riemannian random paths are investigated, including estimation methods, asymptotic properties and an application to brain functional connectivity.

E0529: **Image forecasting using dynamical functional time series models**
Presenter: **Julian Austin**, Newcastle University, United Kingdom
Co-authors: Jian Qing Shi, Zhenhong Li

The advancement of remote sensing technologies over the last few decades has meant a plethora of earth observation data is becoming readily available. Such data is often viewed as imagery and can be used for climate analysis, land monitoring or natural hazard studies. However, the data sets are often limited due to constraints on acquisition times. Forecasting or interpolation of these images would provide additional information which can be fed back into management scenarios. Forecasting of remotely sensed images is often difficult due to the high dimensionality of the data coupled with the spatial and temporal dependency among images. We consider the approach of treating imagery from a functional point of view. We utilise functional decompositions to reduce the dimensionality of the data before forecasting the simpler series of component scores. We compare functional principal component analysis (fPCA) and functional maximal correlation factors (fMAF) methods of decomposition to see the impact of forcing correlation in our functional components on our forecast results. We find that both fPCA and fMAF are capable of producing reasonable results, and forcing temporal correlation in fMAF decomposition causes more stable component scores for forecasting and interpolation.

E0997: **Comparing paired distributions using a functional data approach**
Presenter: **Juhyun Park**, ENSIIE, France
Co-authors: Anne Gegout-Petit

A paired hypothesis testing problem arises when the same variable of interest is measured before and after a treatment is applied. We are interested in the case where we only have access to a summary measure in terms of distribution functions of a common random variable. Hence the treatment effect is assumed to be reflected in the change in the distribution functions. Although the data as distribution functions could be viewed as instances of functional data, the standard testing framework based on the mean change in an L_2 Hilbert space poses several difficulties due to the inherent constraints on the distribution functions. Instead, we propose to measure a nonlinear change by means of a warping function between the paired distribution functions and formulate a one-sample hypothesis testing problem based on the median of the warping functions. In order to properly define a functional median, we equip the space of warping functions with a transformation induced norm and define a hypothesis using a norm-based geometric median. A version of a permutation test and an asymptotic test are developed, and their performance is compared with both simulation studies and a real data example.

EO708 Room R12 RECENT DEVELOPMENTS IN MODEL-BASED CLUSTERING **Chair: Monia Ranalli**

E0542: **Clustering data with nonignorable missingness using semi-parametric mixture models**
Presenter: **Matthieu Marbac**, CREST - ENSAI, France
Co-authors: Marie du Roy de Chaumaray

The focus is on clustering continuous data sets subject to nonignorable missingness. We perform clustering with a specific semi-parametric mixture, avoiding the component distributions and the missingness process to be specified, under the assumption of conditional independence given the component. Estimation is performed by maximizing an extension of smoothed likelihood allowing missingness. This optimization is achieved by a Majorization-Minorization algorithm. We illustrate the relevance of our approach by numerical experiments. Under mild assumptions, we show the identifiability of our model, the monotony of the MM algorithm as well as the consistency of the estimator. We propose an extension of the new method to the case of mixed-type data that we illustrate on a real data set.

E0623: **Model based-clustering via mixtures of two new matrix-variate distributions**
Presenter: **Salvatore Daniele Tomarchio**, University of Catania, Italy
Co-authors: Antonio Punzo, Luca Bagnato

Two matrix-variate distributions, both elliptical heavy-tailed generalization of the matrix-variate normal distribution, are introduced. For the nested matrix-variate normal distribution, their probability density functions are characterized by only one additional parameter that governs the tail-weight. Both distributions are then used for model-based clustering via finite mixture models. Being able to handle data with atypical observations in a better way than the matrix-variate normal mixture, the proposed models can avoid the disruption of the true underlying group structure. Several EM-based algorithms are implemented for parameter estimation and tested in terms of computational times and parameter recovery. Furthermore, these mixture models are fitted to simulated and real data, and their fitting and clustering performances are analyzed and compared to those obtained by other well-established competitors.

E0800: **Non-metric unfolding on augmented data matrix: A copula-based approach**
Presenter: **Marta Nai Ruscone**, Universita degli Studi di Genova, Italy
Co-authors: Antonio Dambrosio

Unfolding applies multidimensional scaling to an off-diagonal $n \times m$ matrix, representing the scores (or the rank) assigned to a set of m items by n individuals or judges. The goal is to obtain two configurations of points representing the position of the judges and the items in a reduced geometrical space. Each point, representing each individual, is considered as an ideal point so that its distances to the object points correspond to the preference scores. Unfolding can be seen as a special case of multidimensional scaling because the off-diagonal matrix is considered as a block of an ideal distance matrix in which both the within judges and the within items dissimilarities are missing. The presence of blocks of missing data causes the phenomenon of the so-called degenerate solutions. To tackle the problem, several methods have been proposed. By following a previous approach, we adopt the strategy of augmenting the data matrix, trying to build a complete dissimilarity matrix, and then applying any MDS algorithms. To augment the data matrix, we use copulas-based association. Both experimental evaluations and applications to well-known real data sets show that the proposed strategy produces non-degenerate non-metric unfolding solutions.

EO213 Room R13 RECENT DEVELOPMENTS IN IMAGING DATA ANALYSIS Chair: Farouk Nathoo

E0345: **Bayesian image analysis in transformed spaces (BITS) and the BIFS/WIMP Python packages**
Presenter: **John Kornak**, University of California, San Francisco, United States
Co-authors: Karl Young

Bayesian image analysis can improve image quality by balancing a priori expectations of image characteristics with a model for the noise process. We will give a reformulation of the conventional image space Bayesian image analysis paradigm into Fourier and wavelet spaces. By specifying the Bayesian model in a transformed space, spatially correlated priors, that are relatively difficult to model and compute in conventional image space, can be efficiently modeled as a set of independent processes in an appropriately transformed space. The originally inter-correlated and high-dimensional problem in image space is thereby broken down into a series of (trivially parallelizable) independent one-dimensional problems. We will describe and show examples of the Bayesian image analysis in transformed space (BITS) modeling approach for both Fourier (BIFS) and wavelet (WIMP) space using both parametric and data-driven priors. In the process, we will showcase our Python package(s): BIFS/WIMP that can allow easy and fast implementation of BITS.

E0931: **An adaptive multilayer basis approach for task fMRI data**
Presenter: **Michelle Miranda**, University of Victoria, Canada
Co-authors: Jeffrey Morris

A novel model is proposed to analyze task fMRI data that learns information in an adaptive way, carefully considering the complex structure of the brain data. The approach provides a sparse spatial representation of the brain while yielding full Bayesian inference at the voxel and ROI level with incredible computational speed. In addition, the proposed model allows for free full Bayesian inference on the residual connectivity, which can help scientist gain insights of the underlying brain function. We incorporate biological information from the brain's Regions of Interest (ROIs), accounting for both local correlation in each ROI and distant correlation between ROIs. Time dependencies in the BOLD time series are also considered by projecting the time course into a wavelet space. We then model the final set of bases by assuming a long memory process that accounts for differences in each wavelet decomposition level. We present a simulation study that shows increased power to detect activation when using the proposed composite-hybrid approach. We apply our method to the Working Memory task data from the Human Connectome Project.

E1100: **Permutation-based inference for spatially localized signals in longitudinal MRI data**
Presenter: **Mark Fiecas**, University of Minnesota, United States
Co-authors: Jun Young Park

Alzheimer's disease is a neurodegenerative disease in which the degree of cortical atrophy in specific structures of the brain serves as a useful imaging biomarker. Recent approaches using linear mixed-effects (LME) models in longitudinal neuroimaging are powerful and flexible in investigating the temporal trajectories of cortical thickness. However, massive-univariate analysis, a simplified approach that obtains a summary statistic (e.g., a p-value) for every vertex along the cortex, is insufficient to model cortical atrophy because it does not account for spatial similarities of the signals in neighboring locations. In this article, we develop a permutation-based inference procedure to detect spatial clusters of vertices showing statistically significant differences in the rates of cortical atrophy. The proposed method, called SpLoc, uses spatial information to combine the signals adaptively across neighboring vertices, yielding high statistical power while controlling family-wise error rate (FWER) accurately. When the global null hypothesis is rejected, we use a cluster selection algorithm to detect the spatial clusters of significant vertices. We validate our method using simulation studies and apply it to the Alzheimer's Disease Neuroimaging Initiative (ADNI) data to show its superior performance over existing methods.

EO359 Room R14 RECENT DEVELOPMENT IN HIGH DIMENSIONAL METHODS Chair: Yang Ni

E0423: **Bayesian inference of causal effects from Gaussian graphical models**
Presenter: **Federico Castelletti**, Universita Cattolica del Sacro Cuore (Milan), Italy

It is assumed that multivariate observational data are generated from a distribution whose conditional independencies are encoded in a Directed Acyclic Graph (DAG). For any given DAG, the causal effect of a variable onto another one can be evaluated through intervention calculus. A DAG is typically not identifiable from observational data alone. However, its Markov equivalence class (a collection of DAGs) can be estimated from the data. As a consequence, for the same intervention, a set of causal effects, one for each DAG in the equivalence class, can be evaluated. We propose a Bayesian methodology which combines structure learning of DAGs and causal effect estimation. As a consequence, our approach fully accounts for the uncertainty around the underlying graphical structure, which is crucial for a correct estimation of the causal effect of an intervention on one variable w.r.t. another. We demonstrate the merits of our method in simulation studies, wherein comparisons with current state-of-the-art procedures turn out to be highly satisfactory. Finally, we examine a real data set of gene expressions for Arabidopsis thaliana.

E0613: **Approximate Laplace approximation**
Presenter: **David Rossell**, Universitat Pompeu Fabra, Spain
Co-authors: Anirban Bhattacharya

Bayesian and L_0 model selection strategies require conducting an integration or maximization exercise for each candidate model, in order to assign model posterior probabilities/score. When the number of models is large, and there is no closed-form expression for the integral/maximized likelihood, such computations can become cumbersome. We present a simple yet powerful idea based on the Laplace approximation (LA) to an integral. LA uses a quadratic Taylor expansion around the mode of the integrand, which typically has good accuracy, but requires optimization. We propose the approximate Laplace approximation (ALA), which uses a Taylor expansion around the null parameter value. ALA brings significant speed-ups by avoiding optimizations altogether, and the sharing of sufficient statistics shared across models. We prove that ALA provides an approximate inference method equipped with strong model selection properties in the family of non-linear GLMs, attaining comparable rates to exact computation. We also show that when the model is misspecified, the ALA rates can actually be faster than for exact computation, depending on the type of misspecification. We illustrate with examples in linear, logistic and survival regression with non-local priors.

E1086: **A Bayesian graphical model for microbiome data**
Presenter: **Hee Cheol Chung**, Texas A&M University, United States
Co-authors: Irina Gaynanova, Yang Ni

Microorganisms present in nature often co-occur, forming communities that have been found to play a critical role in host health. The recent development of high-throughput sequencing technologies provide opportunities for a deeper understanding of microbial communities. However, due to limited sequencing depth, microbiome data have a large excess of technical zeros, which poses a statistical challenge for reverse-engineering microbial association networks. We propose a Bayesian graphical model based on a latent Gaussian copula by modeling technical zeros as censored observations of a latent variable. Microbes' evolutionary information is incorporated as a prior distribution for edge inclusion probabilities using the diffusion process and latent position model. Numerical studies based on simulated examples suggest that the phylogenetic tree prior significantly improves estimation performance. We present an analysis of a quantitative microbiome profiling data set, and compare it to existing methods.

EO445 Room R15 ADVANCES IN THE STATISTICAL ANALYSIS OF DEPENDENT NETWORK DATA Chair: Michael Schweinberger

E0220: **Identifiability and consistency of network inference using the hub model and variants**
Presenter: **Yunpeng Zhao**, Arizona State University, United States
Statistical network analysis primarily focuses on inferring the parameters of an observed network. In many applications, especially in the social sciences, the observed data is the groups formed by individual subjects. In these applications, the network is itself a parameter of a statistical model. A model-based approach, called the hub model, has been recently proposed to infer implicit networks from grouping behavior. The hub model assumes that each member of the group is brought together by a member of the group called the hub. The hub model belongs to the family of Bernoulli mixture models. Identifiability of parameters is a notoriously difficult problem for Bernoulli mixture models. We prove identifiability of the hub model parameters and estimation consistency under mild conditions. Furthermore, we generalize the hub model by introducing a model component that allows hubless groups in which individual nodes spontaneously appear independent of any other individual. We refer to this additional component as the null component. The new model bridges the gap between the hub model and the degenerate case of the mixture model – the Bernoulli product. Identifiability and consistency are also proved for the new model. Numerical studies are provided to demonstrate the theoretical results.

E0369: **Randomization tests for spillovers under general interference: A graph-theoretic approach**
Presenter: **Panagiotis Toulis**, University of Chicago, United States
Co-authors: David Puelz, Avi Feller, Guillaume Basse
Interference exists when the outcome of a unit depends on another units treatment assignment. For example, intensive policing on one street could have a spillover effect on neighboring streets. Classical randomization tests typically break down in this setting because many null hypotheses of interest are no longer sharp under interference. A promising alternative is to instead construct a conditional randomization test on a subset of units and assignments for which a given null hypothesis is sharp. Finding these subsets is challenging, however, and existing methods either have low power or are limited to special cases. We propose valid, powerful, and easy-to-implement randomization tests for a general class of null hypotheses under arbitrary interference between units. The key idea is to represent the hypothesis of interest as a bipartite graph between units and assignments and to find a biclique of this graph. Importantly, the null hypothesis is sharp for the units and assignments in this biclique, enabling randomization-based tests conditional on the biclique. We can apply off-the-shelf graph clustering methods to find such bicliques efficiently and at scale. We illustrate this approach in settings with clustered interference and show advantages over methods designed specifically for that setting. We then apply our method to a large-scale policing experiment in Medellin, Colombia, where interference has a spatial structure.

E0964: **Maximum pseudolikelihood estimation for models of social network data**
Presenter: **Jonathan Stewart**, Florida State University, United States
Co-authors: Michael Schweinberger
The statistical analysis of social network data requires both methods and theory for dependent data in some of the most challenging scenarios. Often, we obtain only a single observation of dependent relationships or interactions in a social network and wish to estimate and infer statistical models with parameter vectors of (possibly) increasing dimension. On the methodological side, the estimation of statistical models for dependent network data gives rise to significant computational challenges. We revisit the familiar maximum pseudolikelihood estimation paradigm and demonstrate how it offers both a scalable and accurate alternative to the gold-standard Monte-Carlo maximum likelihood estimation. On the theoretical side, we demonstrate that many statistical models of social network data possess important conditional independence properties, and discuss how naturally occurring structure in social networks facilitate statistical estimation of complex models. An example of a model for social network data capturing brokerage is given. Theoretical guarantees which highlight the key points are also presented.

EO437 Room R16 RECENT ADVANCES IN THEORY AND METHODS FOR SPATIOTEMPORAL MODELING Chair: Zeda Li

E0924: **Adaptive frequency band analysis for nonstationary functional time series**
Presenter: **Scott Bruce**, George Mason 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 its time-varying power spectrum, which describes the contribution to the variability of a functional time series from waveforms oscillating at different frequencies over time. Practitioners seeking low-dimensional summary measures of the power spectrum often partition frequencies into bands and create collapsed measures of power within these bands. However, standard frequency bands have largely been developed through subjective inspection of time series data and may not provide adequate summary measures of the power spectrum. We provide an adaptive frequency band estimation for nonstationary functional time series that adequately summarizes the time-varying dynamics of the series and simultaneously accounts for the complex interaction between the functional and temporal dependence structures. We develop scan statistics that can be used to detect changes in the frequency domain. We establish theoretical properties of this statistic and develop a computationally-efficient, scalable algorithm for 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.

E1075: **Frequency domain analysis for structural breaks of non-stationary spatial data**
Presenter: **Weiyu Zhou**, George Mason University, United States
Co-authors: Pramita Bagchi
Appropriate modelling of the second-order structure is of prime importance in spatial data analysis. A commonly used assumption for modelling spatial covariance is the assumption of second-order stationarity. However, this assumption is often violated in practice, and its misspecification can lead to a wrong inference. We propose to develop a methodology to test the assumption of stationarity in spatial data and identify stationary (or approximately stationary) sub-regions. We define a frequency domain based spatial process, which takes a high value near any location with second-order change in the underlying random field. We propose a consistent estimator of the aforementioned spatial process based on a partial observation of the random field.We establish the asymptotic convergence of a centered and scaled version of this estimator to Gaussian process for a block-wise stationary random field and propose a consistent level α test for stationarity based on the asymptotic distribution based on the simulated quantiles. We also propose an algorithm to identify the different stationary regions for a block-wise stationary random field or regions with similar second-order property for the locally stationary random field.

E1020: **Random surface covariance estimation by shifted partial tracing**
Presenter: **Tomas Masak**, EPFL, Switzerland
The problem of covariance estimation for replicated surface-valued processes is examined 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 non-parametric 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, our methods are shown to yield consistent estimators even under noisy discrete observation, without the need for smoothing. Further to deriving the convergence

rates and limit theorems, we also show that the implementation of our estimators, including for prediction, comes at no computational overhead relative to a separable model. Finally, we demonstrate empirical performance and computational feasibility of our methods in an extensive simulation study and on a real data set.

EO534 Room R18 STATISTICAL ADVANCES ON MICROBIOME DATA ANALYSIS I Chair: Qiwei Li

E0403: **A Bayesian joint model for microbiome data**
Presenter: **Matthew Koslovsky**, Colorado State University, United States

One of the major research questions regarding human microbiome studies is the feasibility of designing interventions that modulate the composition of the microbiome to promote health and cure disease. This requires an extensive understanding of the modulating factors of the microbiome, such as dietary intake, as well as the relationship between microbial composition and phenotypic outcomes, such as body mass index (BMI). Previous efforts have modeled these data separately, employing two-step approaches that can produce biased interpretations of the results. We present a Bayesian joint model that simultaneously identifies clinical covariates associated with microbial composition data and predicts a phenotypic response using information contained in the compositional data. We apply our model to understand the relations between dietary intake, microbial samples, and BMI. In this analysis, we find numerous associations between microbial taxa and dietary factors that may lead to a microbiome that is generally more hospitable to the development of chronic diseases, such as obesity.

E0474: **Visualization and unsupervised clustering of microbiome data**
Presenter: **Yushu Shi**, The University of Missouri Columbia, United States
Co-authors: Christine Peterson, Liangliang Zhang, Robert Jenq, Kim-Anh Do

Microbiome plays an important role in human health and disease. We will first present aPCoA, an easy-to-use tool available as both an R package and a Shiny app, which improves data visualization by adjusting confounding covariates in a PCoA plot under non-Euclidean distance and enhances the presentation of the effects of interest. Then, we will briefly discuss commonly used metrics in unsupervised clustering of microbiome data and propose a new metric, which combines Bray Curtis and unweighted UniFrac distances and gives better performance on tested datasets. In the end, we will introduce a Bayesian model using Dirichlet tree multinomial mixture to cluster human microbiome data, which captures the tree-based topological structure of microbiome data and informatively selects tree nodes contributing to clustering.

E0579: **Bayesian modeling of metagenomic sequencing data for discovering microbial biomarkers in colorectal cancer**
Presenter: **Shuang Jiang**, Southern Methodist University, United States
Co-authors: Qiwei Li, Xiaowei Zhan, Guanghua Xiao, Andrew Koh

Colorectal cancer (CRC) is a major cause of morbidity and mortality globally. Reductions in mortality can be achieved through the detection and treatment of early-stage CRC patients. Colonoscopy is currently the most effective CRC screening test in nowadays. However, it is costly, invasive, and requires anesthesia. A simple, non-invasive test with high accuracy for CRC is urgently needed. Several recent CRC studies have demonstrated a significant association between tumorigenesis and abnormalities in the microbial community. Those findings shed light on utilizing microbial taxa as noninvasive CRC biomarkers. We propose a Bayesian hierarchical framework to identify a set of differentially abundant taxa, which could potentially serve as microbial biomarkers. The bottom level is a multivariate count generative model that links the observed counts in each sample to their latent normalized abundances. The top-level is a Gaussian mixture model with a feature selection scheme for identifying those taxa whose normalized abundances are discriminatory between different phenotypes. The model further employs Markov random field priors to incorporate taxonomic tree information to identify microbial biomarkers at different taxonomic ranks. A CRC case study demonstrates that a resulting diagnostic model trained by the microbial signatures identified by our model in a CRC cohort can significantly improve the current predictive performance in another independent CRC cohort.

EO606 Room R20 CAUSAL INFERENCE METHODS IN GENETIC STUDIES Chair: Josee Dupuis

E0289: **Bayesian networks with missing data imputation enable exploratory analysis of causal complex biological relationships**
Presenter: **Heather Cordell**, Newcastle University, United Kingdom
Co-authors: Richard Howey

Bayesian networks can be used to identify possible causal relationships between variables based on their conditional dependencies and independencies, particularly in complex scenarios with many measured variables. When there is missing data, the standard approach is to remove every individual with missing data before performing any Bayesian network analysis. This can be wasteful and undesirable when there are many individuals with missing data, perhaps with only one or a few variables missing, motivating the use of imputation. We present a new imputation method designed to increase the power to detect causal relationships, where the data may be a mixture of both discrete and continuous variables. This method uses a version of nearest neighbour imputation, whereby missing data from one individual is replaced with data from another individual, their nearest neighbour. For each individual with missing data, subsets of variables that can be used to find the nearest neighbour are chosen by bootstrapping the complete data to estimate a Bayesian network. We show that this approach leads to marked improvements in recall and precision, and we apply the approach to data from a recent study that investigated the causal relationship between methylation and gene expression in rheumatoid arthritis patients.

E0720: **Accounting for Winner's curse, weak instrument bias and pleiotropy in Mendelian randomization studies**
Presenter: **Jack Bowden**, University of Exeter, United Kingdom

Mendelian randomization (MR) is the science of augmenting the analysis of observational data with genetic information to uncover the causal mechanisms of disease. In an MR analysis, single nucleotide polymorphisms (SNPs) are assumed to be an instrumental variable (IV) for a given exposure trait. In the MR field, research has focused on the development of methods to guard against incorrect inferences when using SNPs that exert direct (or pleiotropic) effects on the outcome, not through the exposure. So far, relatively little attention has been given to the issue of bias due to 'Winner's curse', induced when the same data are used to select a small number of SNPs as IVs from a much larger candidate pool based on a p-value threshold. A simple way of removing Winner's curse is to use separate data sets for SNP discovery and MR model fitting, but this is statistically inefficient and is also more vulnerable to weak instrument bias. We describe a method for combining the SNP-discovery into the MR analysis to deliver efficient and unbiased estimates of a causal effect, whilst simultaneously accounting for weak instrument bias and pleiotropy. Our approach is based on the method of uniform minimum variance conditionally unbiased estimation (UMVCUE), which has been used as a technique for bias adjustment in genome-wide association studies and multi-arm multi-stage trials.

E1070: **MR genius: A principled approach to robust Mendelian randomization inference**
Presenter: **Eric Tchetgen Tchetgen**, The Wharton School, University of Pennsylvania, United States

Mendelian randomization (MR) is a popular instrumental variable (IV) approach, in which one or several genetic markers serve as IVs that can sometimes be leveraged to recover valid inferences about a given exposure-outcome causal association subject to unmeasured confounding. A key IV identification condition known as the exclusion restriction states that the IV cannot have a direct effect on the outcome which is not mediated by the exposure in view. In MR studies, such an assumption requires an unrealistic level of prior knowledge about the mechanism by which genetic markers causally affect the outcome. As a result, possible violation of the exclusion restriction due to pleiotropic genetic effects can seldom be ruled out in practice. To address this concern, we introduce a new class of IV estimators which are robust to violation of the exclusion restriction under data generating mechanisms commonly assumed in MR literature. The proposed approach named "MR G-Estimation under No Interaction

with Unmeasured Selection" (MR GENIUS) improves on Robins' G-estimation by making it robust to both additive unmeasured confounding and violation of the exclusion restriction assumption. Time permitting we will also discuss a many weak invalid IV MR GENIUS approach which appropriately accounts for the fact that in addition to violations due to pleiotropic effects, genetic IVs typically only exhibit a weak effect on phenotypes defining the exposure of interest.

EO636 Room R21 ADVANCES IN CAUSAL SURVIVAL ANALYSIS Chair: Jenny Haggstrom

E0581: **Robust estimation of the average treatment effects in presence of right-censoring and competing risks**
Presenter: **Brice Ozenne**, University of Copenhagen, Denmark
Co-authors: Thomas Scheike, Thomas Alexander Gerds

Average treatment effects (ATE) are important parameters in pharmacoepidemiology where the aim is to evaluate differences between treatments based on health care databases. Estimation of the ATE is complicated by the occurrence of competing events (e.g. death), patient drop-out, and confounders present in non-randomized data. Several types of estimators for the ATE can be derived based on working regression models for the outcome, censoring, and treatment distributions. However, the traditional G-formula or inverse probability weighting (IPW) estimators require well-specified working models to be unbiased. We will present how results from the semi-parametric theory can be used to derive a doubly robust estimator for the ATE. We show, both theoretically and using simulation studies, that the proposed estimator is robust to misspecification of some of the working models and compare it to G-formula and IPW estimators. We will also discuss the use of the functional delta method to obtain the asymptotic distribution of the robust ATE estimator. The proposed robust ATE estimator is implemented in the ate function of the riskRegression package (available on CRAN).

E0599: **Instrumental variable estimation of causal hazard ratio**
Presenter: **Linbo Wang**, University of Toronto, Canada
Co-authors: Eric Tchetgen Tchetgen, Torben Martinussen, Stijn Vansteelandt

Cox's proportional hazards model is one of the most popular statistical models to evaluate associations of a binary exposure with a censored failure time outcome. When confounding factors are not fully observed, the exposure hazard ratio estimated using a Cox model is not causally interpretable. To address this, we propose novel approaches for identification and estimation of the causal hazard ratio in the presence of unmeasured confounding factors. Our approaches are based on a binary instrumental variable and an additional no-interaction assumption. We derive, to the best of our knowledge, the first consistent estimator of the population marginal causal hazard ratio within an instrumental variable framework. Our estimator admits a closed-form representation. Hence it avoids the drawbacks of estimating equation based estimators. The approach is illustrated via simulation studies and data analysis.

E1120: **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.

EO760 Room R22 BAYESIAN DATA INTEGRATION OF COMPLEX OBJECTS Chair: Shuang Zhou

E0518: **Analysis of professional basketball field goal attempts via a Bayesian matrix clustering approach**
Presenter: **Guanyu Hu**, University of Missouri Columbia, United States

A model-based clustering approach for matrix response data is developed to analyze the underlying heterogeneity structure of shot selection among professional basketball players in the NBA. Particularly, we propose a mixture of finite mixtures (MFM) model for heterogeneity learning. Our proposed method estimates the number of clusters and cluster configurations simultaneously. The theoretical properties of our proposed method are established. An efficient Markov Chain Monte Carlo (MCMC) algorithm is designed for our proposed model. Extensive simulation studies are carried out to examine the empirical performance of the proposed methods. We further apply the proposed methodology to analyze shot charts of selected players in the NBAs 2017-2018 regular season.

E0556: **Bayesian Spatial homogeneity pursuit of functional data**
Presenter: **Junxian Geng**, Boehringer Ingelheim, United States
Co-authors: Yishu Xue, Huiyan Sang, Guanyu Hu

An income distribution describes how an entity's total wealth is distributed amongst its population. A problem of interest to regional economics researchers is to understand the spatial homogeneity of income distributions among different regions. In economics, the Lorenz curve is a well-known functional representation of income distribution. We propose a mixture of finite mixtures (MFM) model as well as a Markov random field constrained mixture of finite mixtures (MRFC-MFM)model in the context of spatial functional data analysis to capture spatial homogeneity of Lorenz curves. We design efficient Markov chain Monte Carlo (MCMC) algorithms to simultaneously infer the posterior distributions of the number of clusters and the clustering configuration of spatial functional data. Extensive simulation studies are carried out to show the effectiveness of the proposed methods compared with existing methods. We apply the proposed spatial functional clustering method to state-level income Lorenz curves from the American Community Survey Public Use Microdata Sample (PUMS) data. The results reveal several important clustering patterns of state-level income distributions across the US.

E0667: **Data integration using hierarchical Gaussian process models under shape constraints**
Presenter: **Shuang Zhou**, Arizona State University, United States

Data integration has been a hot topic in real-world applications that combines data residing at different sources and extracts the shared information across sources. Hierarchical Bayesian models are a powerful tool for modelling grouped data by modelling the data and their interaction across the groups via hierarchies. We develop a method for data integration under multiple constraints using a hierarchical constrained regression with basis expansion approach. At the global level, we can incorporate multiple constraints simultaneously by finding a one-to-one mapping of the constraints on the coefficient space from the original function space under a suitable basis. At the group level, we integrate a multiplicative random effect into the sub-models with the knowledge of data annotation to estimate the group-wise unknown response deviation. We apply our model to the proton radius puzzle problem in nuclear physics, where the constraints come from the law of physics and the unknown experimental errors associate with the data sources. We recover both the global parameters and the group errors related to the sources in the synthetic data analysis and in the real application we provide reliable analyses to reconcile with the new results for the proton radius extraction.

EO510 Room R23 STATISTICAL LEARNING FOR DECISION-MAKING SYSTEMS Chair: Matteo Borrotti

E0543: **Design of experiments on networks for decision making**
Presenter: **Vasiliki Koutra**, University of Southampton, United Kingdom
Co-authors: Steven Gilmour, Ben Parker

Design of experiments provides a formal framework for the collection of data to aid decision making, ranging from what drug treatment is most effective through the choice of wheat variety to maximise yield to the selection of an internet advertisement to optimise revenue. When such experiments are performed on connected units, i.e. linked through a network, the resulting design, analysis and decision making is 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 methodology for constructing efficient designs which controls for variation among the experimental units from two sources: blocks and network interference, so that the direct treatment effects can be precisely estimated. We provide evidence that our approach can lead to efficiency gains over conventional designs such as randomised designs that ignore the network structure. We illustrate its usefulness for experiments on networks.

E0321: **Bayesian clustering-based adjacency modelling in disease mapping**
Presenter: **Xueqing Yin**, University of Glasgow, United Kingdom
Co-authors: Duncan Lee, Gary Napier, Craig Anderson

Conditional autoregressive (CAR) models are the most common modelling approaches in disease mapping to quantify the spatial pattern in disease risk across n areal units. In these models the spatial autocorrelation is typically induced by a $(n \times n)$ binary neighbourhood matrix based on the sharing a common border specification, such that spatial correlation is always enforced between geographical neighbours. However, geographically adjacent areas may sometimes exhibit step changes in risk due to factors such as population behaviours and socio-economic deprivation. Therefore, we propose a novel methodology to account for these step changes via a two-stage modelling approach. In stage one we produce a set of candidate neighbourhood matrices via a variety of common clustering methods. In the second stage, an appropriate spatial autocorrelation structure is estimated by estimating the neighbourhood matrix as part of a hierarchical Bayesian spatio model. The proposed model yields improved risk estimation and simultaneously identifies clusters of areas exhibiting elevated or reduced risks. The effectiveness of the methodology is evidenced by a simulation study, and the methodology is motivated by a study of respiratory disease risk in Greater Glasgow, Scotland, in 2016.

E0559: **A hybrid approach in dynamic treatment regimes problems: Multivariate Bayesian machine learning**
Presenter: **Edoardo Ghezzi**, University of Milano-Bicocca, Italy
Co-authors: Matteo Borrotti

A dynamic treatment regime (DTR) is a sequence of decision rules, one per stage of intervention, which aims to adapt a treatment plan to the time-varying state of an individual subject. The Bayesian Machine Learning approach (BML) avoids many of the problems arising from the use of other common methods, such as Q-learning, for identifying optimal DTRs. In problems concerning personalized medicine, it is often plausible to have a large set of variables. In such scenarios, the BML approach might be incalculable due to the singularity of the design matrix and therefore, its unfeasible reversibility. The suggested approach, known as Multivariate Bayesian Machine Learning (MBML), consists of an initial variable selection performed by Spike and Slab priors, in particular the independence slab (i-slab), followed by a generalization of the classic BML approach, which enables its use in a multivariate scenario. Through a simulation study, the MBML approach is compared with Q-learning in various two-stage settings; these settings differ in complexity, as they have several sample sizes and dimensionalities, and in regularity conditions, as they vary each stage's coefficients. The MBML approach has given proof of its reliability by showing results which were more accurate than those of Q-learning in almost every scenario.

EO323 Room R25 ALGORITHMIC FAIRNESS WITH STATISTICAL GUARANTEES Chair: Mohamed Hebiri

E1080: **Projection to fairness in statistical learning**
Presenter: **Thibaut Le Gouic**, Institut Mathematiques de Marseille, France
Co-authors: Jean-Michel Loubes, Philippe Rigollet

In the context of regression, we consider the fundamental question of making an estimator fair while preserving its prediction accuracy as much as possible. To that end, we define its projection to fairness as its closest fair estimator in a sense that reflects prediction accuracy. Our methodology leverages tools from optimal transport to constructing efficiently the projection to fairness of any given estimator as a simple post-processing step. Moreover, our approach precisely quantifies the cost of fairness, measured in terms of prediction accuracy.

E1167: **Fair learning: An optimal transport-based approach**
Presenter: **Paula Gordaliza**, Basque Center for Applied Mathematics, Spain
Co-authors: Eustasio del Barrio, Jean-Michel Loubes, Fabrice Gamboa, Laurent Risser, Philippe Besse

The generalization of applications based on ML models in everyday life and the professional world has been accompanied by concerns about the ethical issues that may arise from the adoption of these technologies. First, we motivate the fairness problem by presenting some comprehensive results from the analysis of the disparate impact on a real dataset. We show that trying to make fair ML models may be a particularly challenging task, especially when the training observations contain bias. Then a review of Mathematics for fairness in ML is given with some novel contributions in the analysis of the price for fairness in regression and classification. We recast the links between fairness and predictability in terms of probability metrics. We analyze repair methods based on mapping conditional distributions to the Wasserstein barycenter and propose a random repair. Secondly, we consider the asymptotic theory of the empirical transportation cost. We provide a CLT for the Wasserstein distance between two empirical distributions with different sizes, for observations on R. In the case $p > 1$, the assumptions are sharp in terms of moments and smoothness. We prove results dealing with the choice of centering constants. We provide a consistent estimate of the asymptotic variance, which enables to build two-sample tests and confidence intervals to certify the similarity between two distributions. These are used to assess a new criterion of dataset fairness in classification.

E1123: **Individual fairness through robustness**
Presenter: **Mikhail Yurochkin**, University of Michigan and IBM Research, United States

An approach is considered to train machine learning systems that are fair in the sense that their performance is invariant under certain perturbations to the features. For example, the performance of a resume screening system should be invariant under changes to the name of the applicant or switching the gender pronouns. We connect this intuitive notion of algorithmic fairness to individual fairness and study how to certify ML algorithms as algorithmically fair. We demonstrate the applicability of our framework to supervised learning of neural networks, gradient boosted decision trees and learning to rank problems. We also discuss extensions to the task of auditing ML systems for individual fairness violations. We demonstrate the effectiveness of our approaches on three machine learning tasks that are susceptible to gender and racial biases.

E0886: **A minimax framework for quantifying risk-fairness trade-off in regression**
Presenter: **Nicolas Schreuder**, CREST (UMR 9194), France
Co-authors: Evgenii Chzhen

A theoretical framework is proposed for the problem of learning a real-valued function which meets fairness requirements. This framework is built upon the notion of α-relative (fairness) improvement of the regression function which we introduce using the theory of optimal transport. Setting

$\alpha = 0$ corresponds to the regression problem under the Demographic Parity constraint, while $\alpha = 1$ corresponds to the classical regression problem without any constraints. For $\alpha \in (0,1)$ the proposed framework allows to continuously interpolate between these two extreme cases and to study partially fair predictors. Within this framework we precisely quantify the cost in risk induced by the introduction of the fairness constraint. We put forward a statistical minimax setup and derive a general problem-dependent lower bound on the risk of any estimator satisfying α-relative improvement constraint. We illustrate our framework on a model of linear regression with Gaussian design and systematic group-dependent bias, deriving matching (up to absolute constants) upper and lower bounds on the minimax risk under the introduced constraint.

EC787 Room R24 CONTRIBUTIONS IN STATISTICAL MODELLING Chair: Subir Ghosh

E0425: **GLMcat: An R package for generalized linear models for categorical responses**
Presenter: **Lorena Leon**, Universite de Montpellier, CIRAD, France
Co-authors: Jean Peyhardi, Catherine Trottier

In statistical modeling, there is a wide variety of regression models for categorical responses. Yet, no software encapsulates all of these models in a standardized format. We introduce and illustrate the utility of GLMcat, the R package we developed to estimate generalized linear models implemented under the unified specification (r,F,Z), where r represents the ratio of probabilities (reference, cumulative, adjacent, or sequential), F the cumulative distribution function for the linkage, and Z the design matrix. We present the properties of the four families of models, which must be investigated when selecting the components r, F, and Z. The functions are user-friendly and fairly intuitive; offering the possibility to choose from a large range of models through a combination (r,F,Z). Through different examples, we compare our package with VGAM and ordinal, two popular packages for implementing GLMs for categorical data.

E0340: **A Hidden Markov model addressing ordinal response for non-decreasing processes**
Presenter: **Lizbeth Naranjo Albarran**, Universidad Nacional Autonoma de Mexico UNAM, Mexico
Co-authors: Carlos Javier Perez Sanchez, Yolanda Campos-Roca

Several investigations have recently considered the use of acoustic parameters extracted from speech recordings as an objective and non-invasive tool to perform diagnosis and monitoring of Parkinson's Disease (PD). Repeated speech recordings were obtained from which several acoustic characteristics were extracted. The objective is to monitor the progression of people with PD in the Hoehn and Yahr scale. A Hidden Markov Model (HMM) addressing the ordinal response with some missing data for monotonic non-decreasing processes is proposed. This model used the strength of the HMM to track the progression of the disease at the same time that handles ordinal response with missingness data and non-decreasing of the stages through the time. The way the model is defined allows the derivation of an efficient MCMC algorithm.

E1194: **Network Hawkes process models for exploring latent hierarchy in social animal interactions**
Presenter: **Owen Ward**, Columbia University, United States
Co-authors: Tian Zheng, Anna Smith

Group-based social dominance hierarchies are of essential interest in animal behavior research. Studies often collect aggressive interaction data observed over time, with researchers interested in understanding how the underlying social hierarchy is established and dynamically evolves. Models that capture such dynamic hierarchy are therefore crucial. Traditional ranking methods summarize interactions across time, relying only on aggregate counts. Instead, we take advantage of the interaction timestamps, proposing a series of network point process models with latent ranks. We carefully design these models to incorporate important characteristics of animal interaction data, including the winner effect, bursting and pair-flip phenomena. Through iteratively constructing and evaluating these models we arrive at the final model, utilising a cohort Markov modulated Hawkes process (C-MMHP), which best characterizes all aforementioned patterns observed in the behavior of mice cohorts. We compare all models under study using simulated and real data. Using statistically developed diagnostic perspectives, we demonstrate that the C-MMHP model outperforms other existing methods and models, recovering the underlying rankings when the ground truth is available, and capturing relevant latent ranking structures that lead to meaningful predictions.

CO618 Room R02 ADVANCES IN FINANCIAL ECONOMETRICS Chair: Helena Veiga

C0232: **Multiplicative non-stationary volatility models with exogenous information**
Presenter: **Cristina Amado**, University of Minho, Portugal

A multiplicative nonstationary volatility model is proposed allowing for nonlinear behaviour driven by exogenous information. The new model extends the time-varying GARCH model by including an additional stochastic variable to allow the conditional variance to change smoothly between regimes. Modelling strategies for the proposed model are developed and they rely on Lagrange multiplier tests. The estimation of the model is simplified by employing maximisation by parts and the asymptotic properties of the proposed estimators are also studied. Finite-sample properties of these procedures and statistical tests are examined by simulation. An empirical application to commodity price data illustrates the functioning of the model in practice.

C0242: **Semiparametric Bayesian forecasting for copula stochastic volatility model**
Presenter: **Audrone Virbickaite**, Colegio Universitario Estudios Financieros, Spain
Co-authors: Martina Danielova Zaharieva, Fabian Goessling

The contribution is twofold. First, a new highly flexible semiparametric copula stochastic volatility model (SCSV) is proposed, which is based on an infinite location-scale mixture at the financial return level, and accounts for the asymmetric volatility persistence by using a copula-based transition at the latent volatility level. Second, we propose a new and highly flexible Bayesian sampling algorithm for nonlinear state space models under nonparametric distributions. The estimation framework combines a particle filtering and smoothing algorithm for the latent process with a Dirichlet process mixture model for the error term of the observable variables. In particular, we overcome the problem of the intractable likelihood of the newly proposed SCSV model and avoid constraining the model by transformations or the need for conjugate distributions. We test our approach for several nested model specifications using simulated data and provide return density forecasts. Finally, we present an application study using financial returns of several stock market indices.

C0928: **Forecasting value-at-risk and expected shortfall: A Bayesian approach**
Presenter: **J Miguel Marin**, University Carlos III, Spain
Co-authors: Helena Veiga

The main aim is to investigate whether the asymmetric response of the volatility plays an important role in forecasting the short term horizon VaR and ES and whether the choice of the model determines these forecasts. There is some evidence in the literature reporting that volatility asymmetries included in GARCH type models improve VaR and ES for short horizons. We extend the literature by modeling five international stock market returns according to six volatility models: three belong to the GARCH family while the other three models are in the SV family. All models estimations, VaR and ES forecasts are obtained with a Markov chain Monte Carlo (MCMC) procedure implemented in the R package Nimble and package doParallel that allows for parallel computing. This is a great advantage given that MCMC methods are known to be time-consuming.

CO397 Room R04 QUANTITATIVE APPROACH TO HIGHER EDUCATION RESEARCH Chair: Svetlana Makarova

C0539: **Journal rankings and publication strategy**
Presenter: **Oleksandr Talavera**, University of Birmingham, United Kingdom
Co-authors: Piotr Spiewanowski

The impact of journal ranking systems on the publication outlet choice is investigated. We analyse the life cycle of 15,555 working papers uploaded by UK-based scholars registered on IDEAS/RePEc repository since 2010. The estimates suggest two main facts. Authors strategically choose outlets to maximize their publication scores. The identification strategy is based on exploiting the change in the British ABS journal ranking in 2015. Working papers written before the 2015 ABS journal ranking change are significantly less likely to be published in ex-post downgraded journals. The effect cannot be attributed to the overall change in the journal quality.

C0884: **Rewarding academic journals publications: Efficient mechanisms and empirical assessment**
Presenter: **Wojciech Charemza**, Vistula University, Poland
Co-authors: Michal Lewandowski, Lukasz Wozny

The efficiency of the mechanism of incentivising publications in academic journals by a system of points (or stars) awarded for quality publications is considered. It is assumed that the research supervision body wants to maximise the expected prestige of the aggregate of academic disciplines. It constructs the reward system in such a way that researchers, who each aim to maximise their reward, maximise, through their publication allocation decisions, the objective function of the research supervision body. The main assumptions are that (i) the marginal probabilities of papers acceptance are known; (ii) the conditional probabilities of acceptance and their research types are known to the researchers, but not to the supervision body and (iii), the types are modelled by beta distribution. Building on principal-agent literature with hidden types, we construct a parsimonious theoretical framework allowing to characterise efficient rewarding mechanisms. The main results are (i) the minimal set of assumptions for monotone journal submission strategies, (ii) conditions for optimal journal categorisation and (iii) monotone comparative statics of the optimal journal categorisation wrt shifts of researchers types/abilities. The model is calibrated to the settings of the reward scheme introduced within the Polish higher education reform in 2018.

C0780: **Relationship of university reputation and popularity: Analysis of Google Trends and QS worldwide university ranking**
Presenter: **Krzysztof Rybinski**, Vistula University Warsaw, Poland
Co-authors: Andrzej Wodecki

When universities prepare and execute long-term development strategies, they face tough choices on how to allocate scarce budgets. A frequent dilemma is whether to support research, that is crucial for the university reputation and its standing in the rankings, or to spend more resources on marketing, teaching quality and possibly sports teams, that will make a university more popular. We analyse how the university reputation and university popularity are related by examining the changes in QS ranking scores (overall score, academic reputation and employer reputation) and Search Volume Indexes (in nine search categories). The preliminary analysis conducted for the 2012 - 2020 period for the top 500 universities revealed that there is a strong and significant positive relationship between the two variables. This relationship is statistically robust in both the OLS and the quantile regressions. It does not hold, however, for the best universities and in the short-run. We also documented that the positive reputation-popularity link was powerful in the first half of the 2010s and ceased to exist in the second half, indicating a possible structural change in Google searches.

CO045 Room R07 MACHINE LEARNING TECHNIQUES IN MACROECONOMICS AND FINANCE Chair: Robinson Kruse-Becher

C0333: **The macroeconomy as a random forest**
Presenter: **Philippe Goulet Coulombe**, University of Pennsylvania, United States

Over the last decades, an impressive amount of nonlinearities have been proposed to reconcile reduced-form macroeconomic models with the data. Many of them boil down to have linear regression coefficients evolving through time: threshold/switching/smooth-transition regression; structural breaks and random walk time-varying parameters. While all of these schemes are reasonably plausible in isolation, we argue that those are much more in agreement with the data if they are combined. To this end, we propose Macroeconomic Random Forests, which adapts the canonical Machine Learning (ML) algorithm to the problem of flexibly modeling evolving parameters in a linear macro equation. The approach exhibits clear forecasting gains over a wide range of alternatives and successfully predicts the drastic 2008 rise in unemployment. The obtained generalized time-varying parameters (GTVPs) are shown to behave differently compared to random walk coefficients by adapting nicely to the problem at hand, whether it is regime-switching behavior or long-run structural change. By dividing the typical ML interpretation burden into looking at each TVP separately, I find that the resulting forecasts are, in fact, quite interpretable. An application to the US Phillips curve reveals it is probably not flattening the way you think.

C0380: **Now- and backcasting initial claims with high-dimensional daily internet search-volume data**
Presenter: **Daniel Borup**, Aarhus University, Denmark
Co-authors: David Rapach, Erik Christian Montes Schutte

A sequence of now- and backcasts of weekly unemployment insurance initial claims (UI) is generated based on a rich trove of daily Google Trends (GT) search-volume data for terms related to unemployment. To harness the information in a high-dimensional set of daily GT terms, we estimate predictive models using machine-learning techniques in a mixed-frequency framework. The sequence of now- and backcasts are made ten days to one day before the release of the UI figure on Thursday of each week. In a simulated out-of-sample exercise, now- and backcasts of weekly UI that incorporate the information in the daily GT terms substantially outperform those based on an autoregressive benchmark model, especially since the advent of the COVID-19 crisis. The improvements in predictive accuracy relative to the autoregressive benchmark generally increase as the now- and backcasts include additional daily GT data, with reductions in root mean squared error of up to approximately 50%. Variable-importance measures reveal that the GT terms become more relevant for predicting UI during the crisis. At the same time, partial-dependence plots indicate that linear specifications are largely adequate for capturing the predictive information in the GT terms. We are in the process of creating a website that will provide updated, real-time now- and backcasts of UI daily.

C0749: **Macroeconomic forecasting with deep factor models**
Presenter: **Simon Lineu Umbach**, FernUniversitaet in Hagen, Germany

In many macroeconomic forecasting applications, factor models are used to cope with large datasets. Variational autoencoders with macroeconomic factor modeling are aligned, and an extension is proposed to adapt this framework for forecasting exercises. Variational autoencoders are well suited for nonlinear dimensionality reduction. They estimate the distribution of the common latent variables by combining a statistical factor model with a purely data-driven neural network approach. It is demonstrated that the resulting deep factor model can be interpreted as a flexible nonlinear extension of the standard factor model. In the empirical part, it is analyzed whether factor models augmented by neural networks can achieve superior forecasting power. The results suggest significant improvements in the forecasting accuracy of four major US macroeconomic time series.

CC804 Room R06 CONTRIBUTIONS IN FORECASTING I **Chair: Valderio Anselmo Reisen**

C0833: **Assessing the reliability of aggregated inflation views in the European commission consumer survey**
Presenter: **Maritta Paloviita**, Bank of Finland, Finland
Co-authors: Ewa Stanislawska, Tomasz Lyziak

Using a novel approach, we assess the reliability of aggregated inflation expectations in the European Commission Consumer Survey by identifying individual responses to qualitative and quantitative survey questions that do not match each other. We examine how inconsistent survey responses affect balance statistics, mean inflation expectations and the assessment of the formation of inflation expectations based on the sticky-information model. The Finnish and Polish data indicate that micro-level inconsistencies neither matter for the aggregated inflation views nor explain inflation overestimation bias displayed in the data. Overall, micro-level inconsistencies do not reduce the reliability of the European Commission Consumer Survey.

C1143: **Correlation models for description of market risk factors**
Presenter: **Lukasz Bielak**, Wroclaw University of Science and Technology, Poland

Mining companies have to prepare potential scenarios for main market risk factors. Regardless of the typical uncertainty related to individual price projections, the main challenge is to properly quantify dependencies/relations among main risk factors and its stability over time. Detailed studies of the risk factors dependency structure and finding proper models reflecting such relationship may enable building more adequate forecasts, especially for stress test scenarios. We concentrate on the relations between mentioned factors and using mathematical/statistical methods. We propose a model that takes under consideration the dependencies between mentioned risk factors.

C0569: **Information extraction from the GDELT database to analyse the European sovereign bond market**
Presenter: **Luca Tiozzo Pezzoli**, JRC European Commission, Italy
Co-authors: Sergio Consoli, Elisa Tosetti

A set of news-based indicators are extracted and used to forecast future behaviour of the sovereign bond yield spread in Italy. We use a big, open-source, news-level database known as Global Database of Events, Language and Tone (GDELT) and extract a large number of variables capturing daily variations in the emotional content of news economic and political events, as well as topics popularity, and use these as proxies for market investor's expectations and behaviour. To this end, we adopt a recurrent network model with feedback connections, known as DeepAR, in combination with a number of approaches for variable reduction, and forecast yield spread at different quantiles. Results show good performance of our methodology for the forecasting of the Italian sovereign bond market using the information extracted from GDELT and a deep Long Short-Term Memory Network opportunely trained and validated with a rolling window approach to best accounting for non-linearities in the data.

CG030 Room R03 CONTRIBUTIONS IN FINANCIAL ECONOMETRICS II **Chair: Richard Luger**

C0522: **Composite likelihood estimation of an autoregressive panel ordered probit model with random effects**
Presenter: **Kerem Tuzcuoglu**, Bank of Canada, Canada

Modeling and estimating autocorrelated discrete data can be challenging. We use an autoregressive panel ordered probit model where the autocorrelation in the latent variable drives the serial correlation in the discrete variable. In such a non-linear model, the presence of a lagged latent variable results in an intractable likelihood containing high-dimensional integrals. To tackle this problem, we use composite likelihoods that involve a much lower order of integration. However, parameter identification becomes problematic since the information employed in lower-dimensional distributions may not be rich enough for identification. Therefore, we characterize types of composite likelihoods that are valid for this model and study conditions under which the parameters can be identified. Moreover, we provide consistency and asymptotic normality results of the pairwise composite likelihood estimator and conduct Monte Carlo studies to assess its finite-sample performances. Finally, we apply our method to analyze credit ratings. The results indicate a significant improvement in the estimated probabilities for rating transitions compared with static models.

C1036: **Estimating time-varying networks with a state-space model**
Presenter: **Sandra Paterlini**, University of Trento, Italy
Co-authors: Shaowen Liu, Massimiliano Caporin

A state-space model (SSM) is proposed to estimate dynamic spatial relationships from time-series data. At each time step, the weight matrix, capturing the latent state, is updated by a spatial autoregressive model. Specifically, we consider two types of SSM: the first one calibrates the spatial model to a multivariate regression. In contrast, the second one updates the spatial matrix by leveraging the maximum likelihood (ML) estimation. Simulation results show that the first model performs robustly for all cases, while the performance of the second model is more sensitive to the state dimension. Then, we estimate the time-varying weight matrices with weekly credit default swap (CDS) data for 16 banks and show that the methods identify communities which are coherent with the country-driven partitions.

C1190: **The impact of the curve-fitting procedure on estimation and testing of term structure models**
Presenter: **Jorge Wolfgang Hansen**, Aarhus University and CREATES and the Danish Finance Institute, Denmark
Co-authors: Bent Jesper Christensen

An empirical analysis of interest rates is generally split into an initial yield fitting step and a subsequent model estimation step. We study the impact of the fitting procedure used on the inference in the estimation step, such as on the fit of a particular dynamic term structure model or tests for arbitrage opportunities. We cast the analysis into the Heath-Jarrow-Morton framework and provide an empirical application to US Treasury coupon bond data. We find that the yield curve used in the fitting step, and, in particular, the consistency between the shape of the yield curve and the dynamic term structure model used in the estimation step, has a significant impact on the conclusions drawn from the analysis.

Sunday 20.12.2020 16:30 - 18:10 Parallel Session K – CFE-CMStatistics

EO339 Room R11 RANDOM OBJECTS: REGRESSION, CLUSTERING AND CHANGE-POINTS Chair: Hans-Georg Mueller

E0353: **Point process regression**
Presenter: **Alvaro Gajardo**, University of California Davis, United States
Co-authors: Hans-Georg Mueller

Point processes in time have a wide range of applications such as the number of COVID-19 confirmed cases in a country or the claims arrival process in insurance, among many others. Due to advances in technology, such samples of point processes are increasingly encountered. They are being recorded along with covariates, which contain intrinsic characteristics related to each realization of the process. A key feature of interest is the local intensity function, which measures the rate of occurrence of events per unit time and allows to explore point process data. We consider a novel non-parametric functional regression approach for replications of Cox processes as responses with vector covariates as predictors, which targets the conditional intensity function through the notion of conditional Frechet means. We apply the method to study the effect of the temperature on the Chicago Divvy bike trips, the demand of yellow taxis according to the day of the week in New York as well as COVID-19 case and death point processes over countries depending on social distancing measures.

E0395: **Frechet change point detection for random objects**
Presenter: **Paromita Dubey**, University of California, Davis, United States
Co-authors: Hans-Georg Mueller

Change-point detection is a challenging problem when the underlying data space is a metric space where one does not have basic algebraic operations like the addition of the data points and the scalar multiplication. We propose a method to infer the presence and location of change points in the distribution of a sequence of independent data taking values in a general metric space. Change points are viewed as locations at which the distribution of the data sequence changes abruptly in terms of either its Frechet mean or Frechet variance or both. The proposed method is based on comparisons of Frechet variances before and after putative change-point locations. First, we will establish that under the null hypothesis of no change-point, the limit distribution of the proposed scan function is the square of a standardized Brownian Bridge. Next, we will show that when a change point exists, (1) the proposed test is consistent under contiguous alternatives and (2) the estimated location of the change-point is consistent. We will illustrate the efficacy of the proposed approach in empirical studies and real data applications with sequences of maternal fertility distributions.

E0397: **Inference on compound Cox processes by means of PCP**
Presenter: **Paula Bouzas**, University of Granada, Spain
Co-authors: Nuria Ruiz-Fuentes

The compound Cox process is characterized by whether its intensity or its mean process, so inference on these stochastic processes is essential. Having observed several sample paths of the counting process, functional data analysis and principal components prediction are powerful techniques to estimate and finally to predict them. This inference also allows predicting most of the statistics (mean, mode, probability of a new occurrence, etc.). Additionally, a goodness-of-fit test can be derived to assess if a new observed sample path follows a given compound Cox process. Taking into account that Cox processes with or without random deletions, simultaneous occurrences or a time-space are particular cases of a compound Cox processes, the inference presented can be applied in a variety of real cases. Some examples will illustrate the results.

E0605: **Clustering on the torus by conformal prediction**
Presenter: **Sungkyu Jung**, Seoul National University, Korea, South
Co-authors: Byung-Won Kim, Kiho Park

Motivated by the analysis of torsion (dihedral) angles in the back-bone of proteins, we investigate clustering of bivariate angular data on the torus [0,2) [0,2). We show that naive adaptations of clustering methods, designed for vector-valued data, to the torus are not satisfactory, and propose a novel clustering approach based on the conformal prediction framework. We construct several prediction sets for toroidal data with guaranteed finite-sample validity, based on a kernel density estimate and bivariate von Mises mixture models. From a prediction set built from a Gaussian approximation of the bivariate von Mises mixture, we propose a data-driven choice for the number of clusters, and present algorithms for automated cluster identification and cluster membership assignment. The proposed prediction sets and clustering approaches are applied to the torsion angles extracted from three strains of coronavirus spike glycoproteins (including SARS-CoV-2, contagious in humans). The analysis reveals a potential difference in the clusters of the SARS-CoV-2 torsion angles, compared to the clusters found in torsion angles from two different strains of coronavirus, contagious in animals.

EO447 Room R12 NEW RESEARCH DIRECTIONS IN FUNCTIONAL DATA ANALYSIS Chair: Alexander Aue

E0367: **Testing the conditional mean independence for functional data**
Presenter: **Xiaofeng Shao**, University of Illinois at Urbana-Champaign, United States
Co-authors: Chung Eun Lee, Xianyang Zhang

A new nonparametric conditional mean independence test is introduced for a response variable Y and a predictor variable X where either or both can be function-valued. Our test is built on a new metric, the so-called functional martingale difference divergence (FMDD), which fully characterizes the conditional mean dependence of Y given X and extends a previous MDD. We define an unbiased estimator of FMDD and obtain its limiting null distribution under mild assumptions. Since the limiting null distribution is not pivotal, we adopt the wild bootstrap method to estimate the critical value and show the consistency of the bootstrap test. Promising finite sample performance is demonstrated via simulations and a real data illustration in comparison with several existing tests.

E0390: **Nonstationary fractionally integrated functional time series**
Presenter: **Han Lin Shang**, Macquarie University, Australia
Co-authors: Degui Li, Peter Robinson

A functional version of fractionally integrated time series is studied, which covers the functional unit root as a special case. The functional time series are projected onto a finite number of sub-spaces; the level of nonstationarity allowed to vary over them. Under regularity conditions, we derive a weak convergence result for the projection of the fractionally integrated functional process onto the asymptotically dominant sub-space, which retains most of the sample information carried by the original functional time series. Through the classic functional principal component analysis of the sample variance operator, we obtain the eigenvalues and eigenfunctions which span a sample version of the dominant sub-space. Furthermore, we introduce a simple ratio criterion to consistently estimate the dimension of the dominant sub-space and use a semiparametric local Whittle method to estimate the memory parameter. Monte-Carlo simulation studies and empirical applications are given to examine the finite-sample performance of the developed techniques.

E0563: **Estimation of the covariance function for partially observed functional data**
Presenter: **Marie-Helene Descary**, University of Quebec in Montreal, Canada
Co-authors: Victor Panaretos

The problem of nonparametric estimation of a covariance function on the unit square is considered given a sample of discretely observed fragments of functional data. When each sample path is only observed on a subinterval of length $\delta < 1$, one has no statistical information on the unknown covariance outside a delta-band around the diagonal. A priori, the problem seems unidentifiable without parametric assumptions, but we nevertheless show that nonparametric estimation is feasible under suitable smoothness and rank conditions on the unknown covariance. This remains true even when the observation is discrete. We give precise deterministic conditions on how fine the observation grid needs to be relative to the rank and fragment length for identifiability to hold. We show that our conditions translate the estimation problem to a low-rank matrix completion problem, and construct a nonparametric estimator in this vein. We illustrate our method by simulation and analysis of real data and provide theory to show the validity of the model.

E0735: **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) data-sets which commonly result as high-resolution measurements of stochastic processes varying over a continuum, we propose novel testing procedures 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 the 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 from these novel features, the properties of the tests are robust to any choice of frequency band also enabling 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.

EO596 Room R13 CLIMATE EXTREMES AND DEPENDENCE MODELING Chair: Anna Kiriliouk

E0313: **Modelling non-stationarity in bivariate hazard curves**
Presenter: **Callum Barltrop**, Lancaster University, United Kingdom

Multivariate hazard curves are defined, for a given probability p, as all values of a multivariate random variable for which the joint survival probability is equal to p. For particularly small probabilities, these curves can be used to assess the risk of extreme multivariate events and are often considered to be the natural multivariate extension to a return level. Furthermore, for applications where the risk from combinations of two (or more) variables is considered important, these curves may allow resources to be better allocated. However, difficulties arise when considering environmental data (temperature, wind speed, etc.) since such processes often exhibit non-stationarity. This feature means the underlying hazard curves will also be non-stationary, varying following underlying physical processes and potentially other external factors. We show how to capture non-stationarity in both the margins and dependence structures of bivariate datasets using a previous extension of the model. We then apply this theory to obtain estimates of non-stationary bivariate hazard curves and illustrate the effectiveness of our approach using a simulation study. Moreover, we further demonstrate our approach using data from the 2018 UK Climate Projections (UKCP18) and discuss potential avenues for future research.

E0493: **Space-time Pareto processes to generate scenarios for natural disasters**
Presenter: **Fatima Palacios-Rodriguez**, Universidad Complutense de Madrid, Spain
Co-authors: Gwladys Toulemonde, Julie Carreau, Thomas Opitz

Extreme events of natural phenomena not only can entail material damages but can also have devastating consequences for human societies and ecosystems. In order to assess the risks associated with destructive natural disasters, impact models can be fed with simulations of extreme scenarios for studying the sensitivity to temporal and spatial variability, but available data may be insufficient due to the sparse occurrence of extreme episodes. We introduce a methodology to simulate realistic spatio-temporal extreme fields stochastically. To this end, we use a moderate number of observed extreme space-time episodes to generate an unlimited number of extreme scenarios of any magnitude. The theoretical justification of our framework comes from the extreme-value theory, where generalized Pareto limit processes arise for threshold exceedances. For implementation and illustration of our methodology, we obtain extreme event simulations from hourly gridded precipitation data in Mediterranean France.

E0911: **Spatiotemporal wildfire modeling through point processes with extreme marks**
Presenter: **Jonathan Koh**, EPFL, Switzerland
Co-authors: Thomas Opitz

Accurate modeling of wildfires is essential to gain a better understanding of the mechanisms driving fire-prone ecosystems and to improve risk management. Here, we consider daily summer wildfire records in the French Mediterranean basin during the period 1995–2018. We jointly model the occurrence intensity and the wildfire sizes by combining extreme-value theory and point process tools within a Bayesian hierarchical modelling framework. In the occurrence component, the wildfire ignition locations and times are modelled as a spatiotemporal point pattern generated by a log-Gaussian Cox process. We use the concept of thinning a point process to model the points associated with extreme fires exceeding a high threshold of burnt area. For the size component, we consider the burnt areas as numerical marks for the points and define two subcomponents to model extreme and non-extreme fires, respectively. We capture non-linear relationships between important covariates, such as weather conditions and forest cover, and the different aspects of fire risk, by incorporating component-specific smooth functions that capture seasonal variation. To reveal common effects driving different aspects of wildfire activity, we share latent effects between different components and highlight how this improves interpretability, parsimony and prediction. To achieve tractable inference in our setting with millions of observations, we propose a stratified subsampling scheme that limits information loss.

E1061: **Climate extreme event attribution using multivariate peaks-over-thresholds modeling and counterfactual theory**
Presenter: **Anna Kiriliouk**, University of Namur, Belgium
Co-authors: Philippe Naveau

Numerical climate models are complex and combine a large number of physical processes. They are key tools in quantifying the relative contribution of potential anthropogenic causes (e.g., the current increase in greenhouse gases) on high impact atmospheric variables like heavy rainfall. These so-called climate extreme event attribution problems are particularly challenging in a multivariate context, that is, when the atmospheric variables are measured on a possibly high-dimensional grid. We leverage two statistical theories to assess causality in the context of multivariate extreme event attribution. As we consider an event to be extreme when at least one of the components of the vector of interest is large, extreme-value theory justifies, in an asymptotical sense, a multivariate generalized Pareto distribution to model joint extremes. Under this class of distributions, we derive and study probabilities of necessary and sufficient causation as defined by the counterfactual theory of Pearl. To increase causal evidence, we propose a dimension reduction strategy based on the optimal linear projection that maximizes such causation probabilities. Our approach is tested on simulated examples and applied to weekly winter maxima precipitation outputs of the French CNRM from the recent CMIP6 experiment.

EO496 Room R14 PUBLIC POLICY ANALYSIS AND MACHINE LEARNING II Chair: Michela Bia

E0621: **Entropy balancing for continuous treatments**
Presenter: **Stefan Tuebbicke**, Institute for Employment Research (IAB), Germany
Interest in evaluating the effects of continuous treatments has been on the rise recently. To facilitate the estimation of causal effects in this setting, the present paper introduces entropy balancing for continuous treatments (EBCT) by extending the original entropy balancing for binary treatments. In order to estimate balancing weights, the proposed approach solves a globally convex constrained optimization problem, allowing for much more computationally efficient implementation compared to other available methods. EBCT weights reliably eradicate Pearson correlations between covariates and the continuous treatment variable. This is the case even when other methods based on the generalized propensity score tend to yield insufficient balance due to strong selection into different treatment intensities. Moreover, the optimization procedure is more successful in avoiding extreme weights attached to a single unit. Extensive Monte-Carlo simulations show that treatment effect estimates using EBCT display similar or lower bias and uniformly lower root mean squared error. These properties make EBCT an attractive method for the evaluation of continuous treatments. Software implementation is available for Stata and R.

E0742: **Heterogeneous treatment and spillover effects under clustered network interference**
Presenter: **Falco Joannes Bargagli Stoffi**, Harvard University, United States
Co-authors: Costanza Tortu, Laura Forastiere
The bulk of causal inference studies rules out the presence of interference between units. However, in many real-world settings, units are interconnected by social, physical or virtual ties, and the effect of a treatment can spill from one unit to other connected individuals in the network. In these settings, different people might respond differently not only to the treatment received but also to the treatment received by their network contacts. Understanding the heterogeneity of treatment and spillover effects can help policy-makers in the scale-up phase of the intervention, it can guide the design of targeting strategies with the ultimate goal of making the interventions more cost-effective, and it might even allow generalizing the level of treatment spillover effects in other populations. We develop a machine learning method that makes use of tree-based algorithms and an Horvitz-Thompson estimator to assess the heterogeneity of treatment and spillover effects with respect to individual, neighborhood and network characteristics in the context of clustered network interference. We illustrate how the proposed binary tree methodology performs in a Monte Carlo simulation study. Additionally, we provide an application on a randomized experiment aimed at assessing the heterogeneous effects of information sessions on the uptake of a new weather insurance policy in rural China.

E0696: **On youth training and better job quality: evidence from Job Corps**
Presenter: **German Blanco**, Illinois State University, United States
Co-authors: Alfonso Flores-Lagunes
Most of the economics literature uses monetary compensation as a proxy for job quality. Although active labor market programs generally aim to improve the future quality of life of participants, the evaluation focuses on employment and earnings. We analyze the causal effect of a U.S. job training program for youth—Job Corps (JC)—on future job quality. We define job quality as a linear index that reduces a vector of job characteristics to a scalar quantity. Our index is consistent with the view that workers evaluate a job as a bundle of attributes. Since our index is continuous, we evaluate the distributional impacts of Job Corps training. Given that job quality is defined only for the employed, we address the selection problem by estimating nonparametric bounds on the effects of JC participation for the latent group of individuals that comply with their treatment assignment and would be employed regardless of training. We find that JC has substantial effects on the average quality of jobs attained that are bounded between 14% and 36% of a standard deviation in our job quality index. The distributional analysis suggests that the effects are heterogeneous over the distribution of the job quality index. We also document that females and older participants appear to experience stronger effects relative to males and younger participants, and that this may be driven by experiencing greater access to certain fringe benefits (e.g., flexible work hours and child care).

E0741: **Double machine learning for (weighted) dynamic treatment effects**
Presenter: **Lukas Laffers**, Matej Bel University, Slovakia
Co-authors: Hugo Bodory, Martin Huber
The focus is on evaluating the causal effects of dynamic treatments, i.e. of multiple treatment sequences in various periods, based on double machine learning to control for observed, time-varying covariates in a data-driven way under a selection-on-observables assumption. To this end, we make use of so-called Neyman-orthogonal score functions, which imply the robustness of treatment effect estimation to moderate misspecifications of the dynamic outcome and treatment models. This robustness property permits approximating outcome and treatment models by double machine learning even under high dimensional covariates and is combined with data splitting to prevent overfitting. In addition to effect estimation for the total population, we consider weighted estimation that permits assessing dynamic treatment effects in specific subgroups, e.g. among those treated in the first treatment period. We demonstrate that the estimators are asymptotically normal and root-n consistent under specific regularity conditions and investigate their finite sample properties in a simulation study. Finally, we apply the methods to the Job Corps study to assess different sequences of training programs under a large set of covariates.

EO239 Room R15 RECENT ADVANCES IN MULTIPLE HYPOTHESES TESTING Chair: Shinjini Nandi

E0281: **Fast and powerful conditional randomization testing via distillation**
Presenter: **Lucas Janson**, Harvard University, United States
Given a response Y and covariates (X,Z), we consider testing the null hypothesis that Y is conditionally independent of X given Z. The conditional randomization test (CRT) was recently proposed as a way to use distributional information about $X|Z$ to exactly control Type-I error using any test statistic in any dimensionality without assuming anything about $Y|(X,Z)$. This flexibility in principle allows one to derive powerful test statistics from complex state-of-the-art machine learning algorithms while maintaining statistical validity. Yet the direct use of such advanced test statistics in the CRT is prohibitively computationally expensive, especially with multiple testing, due to the CRT's requirement to recompute the test statistic many times on resampled data. We propose the distilled CRT, a novel approach to using state-of-the-art machine learning algorithms in the CRT while drastically reducing the number of times those algorithms need to be run, thereby taking advantage of their power and the CRT's statistical guarantees without suffering the usual computational expense. Indeed, we show in simulations that all our proposals combined lead to a test that has similar power to the CRT but requires orders of magnitude less computation, making it a practical tool even for large data sets. We demonstrate these benefits on a breast cancer dataset by identifying biomarkers related to cancer stage.

E0485: **Controlling false discovery rate using Gaussian mirrors**
Presenter: **Zhigen Zhao**, Temple University, United States
Co-authors: Xin Xing, Jun Liu
Simultaneously finding multiple influential variables and controlling the false discovery rate (FDR) for linear regression models is a fundamental problem with a long history. We propose the Gaussian Mirror (GM) method, which creates for each predictor variable a pair of mirror variables by adding and subtracting a randomly generated Gaussian random variable and proceeds with a certain regression method, such as the ordinary least-square or the Lasso. The mirror variables naturally lead to a test statistic highly effective for controlling the FDR. Under a weak dependence assumption, we show that the FDR can be controlled at a user-specified level asymptotically. It is shown that the GM method is more powerful than

many existing methods in selecting important variables, subject to the control of FDR, especially under the case when high correlations among the covariates exist.

E0593: **On the development of local FDR-based approach to testing two-way classified hypotheses**
Presenter: **Sanat Sarkar**, Temple University, United States
Co-authors: Shinjini Nandi

Multiple testing of two-way classified hypotheses controlling false discoveries is a commonly encountered statistical problem in modern scientific research. Nevertheless, research focused on developing local FDR (Lfdr) based methods efficiently accommodating such structural information has not yet taken place beyond the one-way classification setting. The first step toward that wider domain is taken. The two-component mixture model is extended from unclassified to two-way classified hypotheses capturing the underlying structure of the hypotheses. The extension provides the foundational framework for the development of newer and potentially powerful Lfdr based multiple testing procedures in their oracle and data-adaptive forms for two-way classified hypotheses.

E0657: **A locally adaptive weighting and screening approach to spatial multiple testing**
Presenter: **Wenguang Sun**, University of Southern California, United States

Exploiting spatial patterns promises to improve both power and interpretability of false discovery rate (FDR) analyses. A new class of locally adaptive weighting and screening (LAWS) rules is developed that directly incorporates useful local patterns into inference by constructing robust and structure-adaptive weights according to the estimated local sparsity levels. LAWS provides a unified framework for a broad range of spatial problems and is fully data-driven. It is shown that LAWS controls the FDR asymptotically under mild conditions on dependence. The finite sample performance is investigated using simulated data, which demonstrates that LAWS controls the FDR and outperforms existing methods in power. The efficiency gain is substantial in many settings. We further illustrate the merits of LAWS through applications to the analysis of 2D and 3D images.

EO464 Room R16 SPATIAL STATISTICS Chair: Soutir Bandyopadhyay

E0316: **Whittle likelihood for irregularly spaced spatial data**
Presenter: **Soutir Bandyopadhyay**, Colorado School of Mines, United States

Under some regularity conditions, including that the process is Gaussian, the sampling region is rectangular, and that the parameter space Θ is compact, it has been shown that the Whittle estimator $\widehat{\theta}_n$ minimizing their version of Whittle likelihood is consistent (for $d \leq 3$). One can construct large sample confidence regions for covariance parameters θ using the asymptotic normality of the Whittle estimator $\widehat{\theta}_n$. However, this requires to estimate the asymptotic covariance matrix, which involves integrals of the spatial sampling density. Moreover, nonparametric estimation of the quantities in the asymptotic covariance matrix requires specification of a smoothing parameter and is subject to the curse of dimensionality. In comparison, we propose a spatial frequency domain empirical likelihood-based approach which can be employed to produce asymptotically valid confidence regions and tests on θ, without requiring explicit estimation of such quantities.

E0399: **Locally scale invariant proper scoring rules**
Presenter: **David Bolin**, King Abdullah University of Science and Technology (KAUST), Saudi Arabia

Averages of proper scoring rules are often used to rank probabilistic forecasts. In many cases, the variance of the individual observations and their predictive distributions vary in these averages. We show that some of the most popular proper scoring rules, such as the continuous ranked probability score (CRPS) which is the go-to score for continuous observation ensemble forecasts, up-weight observations with large uncertainty which can lead to unintuitive rankings. To describe this issue, we define the concept of local scale invariance for scoring rules. A new class of generalized proper kernel scoring rules is derived, and as a member of this class, we propose the scaled CRPS (SCRPS). This new proper scoring rule is locally scale-invariant and therefore works in the case of varying uncertainty. Like CRPS it is computationally available for output from ensemble forecasts and does not require the ability to evaluate the density of the forecast. The theoretical findings are illustrated in a few different applications, where we in particular focus on models in spatial statistics.

E1003: **Machine learning approaches to predicting bacterial infection region of origin from DNA data**
Presenter: **Jordan Taylor**, University of Bath, United Kingdom

Given that an individual has contracted a virus, it is of interest to predict corresponding meta-data from the counts of DNA sub-sequences. Such meta-data target sought is the virus origin where the current technique in the field is to build the phylogenetic tree using the sub-sequences, and by comparing a similarity measure between points, one can infer which region of the world ones virus strain originated from. Instead of building this computationally expensive phylogenetic tree, we instead treat this as a supervised learning problem. We use machine learning techniques to approximate the functional relationship between the observed sequences and the corresponding region location. The problem is then a classification problem with sparse unstructured observations and an imbalanced corresponding set of target region locations. We present initial results using a mix of non-parametric statistical techniques and machine learning methods.

E1017: **Most likely pathways in the ocean**
Presenter: **Adam Sykulski**, Lancaster University, United Kingdom
Co-authors: Michael OMalley

Spatial statistics for ocean data is a rapidly growing area of research. We provide a methodology for estimating the most likely path taken by a particle between two fixed locations on the global ocean surface. Such pathways are useful for understanding ocean circulation in general, and the movement of ocean-borne objects such as plankton, plastic, oil, and debris. Our methodology is purely data-driven using data from GPS-tracked ocean buoys from the Global Drifter Program. We use this data to construct Markov transition matrices and apply Dijkstra's algorithm to find the most likely paths. The novelty is that we apply hexagonal tessellation of the ocean using Uber's H3 index (which we show is far superior to the standard practice of rectangular or lat-lon gridding). We provide techniques for measuring uncertainty by bootstrapping and applying rotations to the hexagonal grid. We also develop a novel method for calculating the estimated travel time associated with a most likely path, which provides useful maps of global ocean connectivity.

EO287 Room R17 METHODS ON HIGH DIMENSIONAL STATISTICS Chair: Andreas Artemiou

E0608: **Graph informed sliced inverse regression**
Presenter: **Eugen Pircalabelu**, Universita catholique de Louvain, Belgium
Co-authors: Andreas Artemiou

A new method is considered for performing dimension reduction when probabilistic graphical models are being used to perform the estimation of parameters. The procedure enriches the domain of application of dimension reduction techniques to settings where (i) the number of variables p in the model is much larger than the available sample size n, (ii) p is much larger than the number of slices H the model uses. The number of projection vectors D can be larger than n. The methodology is developed for the case of the sliced inverse regression model. Still, extensions to other dimension reduction techniques such as sliced average variance estimation or other methods are straightforward. The application on simulated data

reveals that there is a substantial gain to be made by using the graph informed versions even for low dimensional settings. Theoretical derivations and algorithmic implementations are also illustrated.

E0726: **Conditional graphical modeling of multivariate functional data**
Presenter: **Kuang-Yao Lee**, Temple University, United States
Co-authors: Dingjue Ji, Lexin Li, Todd Constable, Hongyu Zhao

Graphical modeling of multivariate functional data is becoming increasingly important in a wide variety of applications. Most existing methods focus on estimating the graph by aggregating samples, but largely ignore the subject-level heterogeneity, which can often be attributed to some external variables. We introduce a conditional graphical model for multivariate random functions, where we treat the external variables as conditioning set and allow the graph structure to vary with the external variables. Our method is built on two new linear operators, the conditional precision operator and the conditional partial correlation operator, which extend the precision matrix and the partial correlation matrix to both the conditional and functional settings. We show their nonzero elements can be used to characterize the conditional graphs, and develop the corresponding estimators. We establish the uniform convergence of the proposed estimators and the consistency of the estimated graph. At the same time, we allow the graph size to grow with the sample size and accommodate both completely and partially observed data. We demonstrate the efficacy of the method through both simulations and a study of brain functional connectivity network.

E0454: **BIG-SIR: a Sliced Inverse Regression approach for massive data**
Presenter: **Benoit Liquet**, Macquarie University and University of Pau and Pays de LAdour, Australia
Co-authors: Jerome Saracco

In a massive data setting, the focus is on a semiparametric regression model involving a real dependent variable Y and a p-dimensional covariable X. This model includes a dimension reduction of X via an index $X'\beta$. The Effective Dimension Reduction (EDR) direction β cannot be directly estimated by the Sliced Inverse Regression (SIR) method due to the large volume of the data. To deal with the main challenges of analysing massive datasets which are the storage and computational efficiency, we propose a new SIR estimator of the EDR direction by following the "divide and conquer" strategy. The data is divided into subsets. EDR directions are estimated in each subset which is a small dataset. The recombination step is based on the optimisation of a criterion which assesses the proximity between the EDR directions of each subset. Computations are run in parallel with no communication among them. A simulation study using our `edrGraphicalTools` R package shows that our approach enables us to reduce the computation time and conquer the memory constraint problem posed by massive datasets. A combination of `foreach` and `bigmemory` R packages are exploited to offer efficiency of execution in both speed and memory. Results are visualised using the bin-summarise-smooth approach through the `bigvis` R package. Finally, we illustrate our proposed approach on a massive airline data set.

EO536 Room R18 STATISTICAL ADVANCES ON MICROBIOME DATA ANALYSIS II Chair: Qiwei Li

E0523: **Estimating microbial interaction networks based on microbiome compositional data**
Presenter: **Hongmei Jiang**, Northwestern University, United States

Microbiome data based on 16S rRNA sequencing are usually summarized as relative abundances in a compositional fashion due to varying sampling/sequencing depths from one sample to another. Characterizing microbial interactions can give us insights into how the microorganisms live and work together as a community in a particular environment. We propose a novel method to estimate multiple different but related microbial interaction networks for high dimensional compositional data across multiple classes. The performance of the proposed method will be investigated by thorough simulation studies and applications to real datasets.

E0604: **Spatial point process models for multivariate microbiome image data**
Presenter: **Kyu Ha Lee**, Harvard T.H. Chan School of Public Health, United States
Co-authors: Brent Coull, Gary Borisy, Floyd Dewhirst, Jessica Mark Welch, Jacqueline Starr

The spatial distribution of microbes is investigated to understand the role of biofilms in human and environmental health. Advances in spectral imaging technologies enable us to display how different taxa (e.g. species or genera) are located relative to one another and host cells. However, most commonly used quantitative methods are limited to describing spatial patterns of bivariate data. Therefore, we propose a flexible multivariate spatial point process model that can quantify spatial relationships among the multiple taxa observable in biofilm images. We have developed an efficient computational scheme based on the Hamiltonian Monte Carlo algorithm, implemented in the R package. We applied the proposed model to tongue biofilm image data.

E0915: **Logistic normal multinominal biclustering mixture model for microbiome count data**
Presenter: **Wangshu Tu**, Binghamton University, United States
Co-authors: Sanjeena Dang, Yuan Fang

The human microbiome plays an important role in human health and disease status. Using next-generation sequencing technologies, it is possible to quantify microbiome composition. Clustering microbiome data can provide valuable information by identifying underlying patterns across samples as well as between the microbes. We develop a novel family of mixtures of logistic normal multinomial models using a modified factor analyzer structure to cluster both the samples and taxa simultaneously. Parameter estimation is done using a variational variant of the alternating conditional expectation conditional maximization(AECM) algorithm that utilizes a variational Gaussian approximation. The proposed method will be illustrated using simulated and real datasets.

E1162: **Discussion for sessionn EO534 and EO536**
Presenter: **Xiaowei Zhan**, The University of Texas Southwestern Medical Center, United States

This discussion will build on the presentations in these sessions and introduce the challenges in statistics and bioinformatics for microbiome data analysis. These will be helpful to bridge the understanding between statisticians and biologists.

EO153 Room R19 SKETCHING AND RELATED METHODS IN REGRESSION Chair: Keith Knight

E0174: **How to reduce dimension with PCA and random projections**
Presenter: **Edgar Dobriban**, University of Pennsylvania, United States
Co-authors: Fan Yang, Sifan Liu, David Woodruff

The aim is to study how to combine "data-oblivious" methods, such as random projections and sketching, and "data-aware" methods, such as principal component analysis (PCA) to get the best of both. We study "sketch and solve" methods that take a random projection (or sketch) first, and compute PCA after. We compute the performance of several popular sketching methods (random iid projections, random sampling, subsampled Hadamard transform, count sketch, etc) in a general "signal-plus-noise" (or spiked) data model. Compared to well-known works, our results (1) give asymptotically exact results, and (2) apply when the signal components are only slightly above the noise, but the projection dimension is non-negligible. We also study stronger signals allowing more general covariance structures. We find that (a) signal strength decreases under projection in a delicate way depending on the structure of the data and the sketching method, (b) orthogonal projections are more accurate, (c) randomization does not hurt too much, due to concentration of measure, (d) count sketch can be improved by a normalization method. The results have implications for statistical learning and data analysis. We also illustrate that the results are highly accurate in simulations and in analyzing empirical data.

E0175: **An econometric perspective on algorithmic subsampling**
Presenter: **Sokbae Lee**, Columbia University, United States
Co-authors: Serena Ng

Datasets that are terabytes in size are increasingly common, but computer bottlenecks often frustrate a complete analysis of the data. While more data are better than less, diminishing returns suggest that we may not need terabytes of data to estimate a parameter or test a hypothesis. But which rows of data should we analyze, and might an arbitrary subset of rows preserve the features of the original data? A line of work is reviewed that is grounded in theoretical computer science and numerical linear algebra, and which finds that an algorithmically desirable sketch, which is a randomly chosen subset of the data, must preserve the eigenstructure of the data, a property known as a subspace embedding. Building on this, we study how prediction and inference can be affected by data sketching within a linear regression setup. We show that the sketching error is small compared to the sample size effect which a researcher can control. As a sketch size that is algorithmically optimal may not be suitable for prediction and inference, we use statistical arguments to provide 'inference conscious' guides to the sketch size. When appropriately implemented, an estimator that pools over different sketches can be nearly as efficient as the infeasible one using the full sample.

E0433: **Compressed and penalized linear regression**
Presenter: **Daniel McDonald**, University of British Columbia, Canada

Modern applications require methods that are computationally feasible on large datasets while retaining good statistical properties. Recent work has focused on developing fast and randomized approximations for solving least squares problems when the data are too large to fit into memory easily or when computations are at a premium. Many of these techniques rely on data-driven subsampling or random compression. We provide new approximate algorithms for solving penalized least-squares problems which have improved statistical performance relative to existing methods. We provide the first efficient methods for tuning parameter selection, compare our methods with current approaches via simulation and application, and provide theoretical intuition which makes explicit the impact of approximation on statistical efficiency and demonstrates the necessity of careful parameter tuning.

E0758: **Shrinkage estimation, model averaging, and degrees of freedom**
Presenter: **Keith Knight**, University of Toronto, Canada

The ridge regression estimate has been shown to be a weighted sum of ordinary least-squares estimates based on subsets of predictors. We define the notion of a "spectrum" of the ridge regression estimate and extend this notion to other linear shrinkage estimates. Applications to ensemble estimation and approximating principal component regression are given.

EO103 Room R20 ADVANCES IN CAUSAL INFERENCE Chair: Michael Daniels

E0208: **Doubly robust nonparametric instrumental variable estimators for censored outcomes**
Presenter: **Nandita Mitra**, University of Pennsylvania, United States
Co-authors: Edward Kennedy, Youjin Lee

Instrumental variable (IV) methods allow us the opportunity to address unmeasured confounding in observational studies and randomized studies with noncompliance. However, there are very few IV methods for censored survival outcomes. We propose nonparametric estimators for the local average treatment effect on survival probabilities under both nonignorable and ignorable censoring. We provide an efficient influence function-based estimator and a simple estimation procedure when the IV is either binary or continuous. The proposed estimators possess double-robustness properties and can easily incorporate nonparametric estimation using machine learning tools. In simulation studies, we demonstrate the flexibility and efficiency of our proposed estimators under various plausible scenarios. We apply our method to the Prostate, Lung, Colorectal, and Ovarian Cancer Screening Trial to estimate the causal effect of screening on survival probabilities and estimate causal contrasts between two interventions under different censoring assumptions.

E0362: **Calibrated estimation of inverse probability of treatment weights for marginal structural models**
Presenter: **Li Su**, University of Cambridge, United Kingdom
Co-authors: Sean Yiu

Marginal structural models (MSMs) with inverse probability-weighted estimators (IPWEs) are widely used to estimate causal effects of treatment sequences on longitudinal outcomes in the presence of time-varying confounding. However, IPWEs for MSMs can be inefficient and unstable if weights are estimated by maximum likelihood. To improve the performance of IPWEs, covariate balancing weight (CBW) methods have been proposed and recently extended to MSMs. However, existing CBW methods for MSMs are inflexible for practical use because they often do not handle non-binary treatments and longitudinal outcomes (instead of eventual outcomes at a study end). We propose a calibration approach to CBW estimation for MSMs that can accommodate (1) both binary and non-binary treatments, (2) eventual and longitudinal outcomes. We develop novel calibration restrictions by eliminating covariate associations with treatment assignment after weighting the observed data sample (i.e., to optimize covariate balance in finite samples). Two different methods are proposed to implement the calibration. We apply our method to a natural history study of HIV for estimating the effects of highly active antiretroviral therapy on CD4 cell counts over time. Extention to handle dependent censoring is also available.

E0595: **Causal inference in spatio-temporal settings**
Presenter: **Georgia Papadogeorgou**, University of Florida, United States
Co-authors: Kosuke Imai, Jason Lyall, Fan Li

Many causal processes have spatial and temporal dimensions. Yet the classic causal inference framework is not directly applicable when the treatment and outcome variables are generated by spatio-temporal processes with an infinite number of possible event locations at each point in time. We take up the challenge of extending the potential outcomes framework and mediation analysis to these settings by formulating the treatment point process as a stochastic intervention. We develop an estimation technique that applies the inverse probability of treatment weighting method to spatially-smoothed outcome surfaces. We demonstrate that the proposed estimator is consistent and asymptotically normal as the number of time period approaches infinity. A primary advantage of our methodology is its ability to avoid structural assumptions about spatial spillover and temporal carryover effects. We use the proposed methods to estimate the effects of American airstrikes on insurgent violence in Iraq.

E0664: **Reinforced designs for observational studies: Multiple instruments plus control groups as evidence factors**
Presenter: **Bikram Karmakar**, University of Florida, United States
Co-authors: Dylan Small, Paul Rosenbaum

Absent randomization inference about the effects caused by treatments depends upon assumptions that can be difficult or impossible to verify. Causal conclusions gain strength from a demonstration that they are insensitive to moderate violations of those assumptions, especially if that happens in each of several statistically independent analyses that depend upon very different assumptions; i.e. if several evidence factors concur. These issues often arise when the investigator has several possible instruments, together with the option of a direct comparison of treated and control subjects. Does each purported instrument satisfy the stringent assumptions required of an instrument? Is a direct comparison without instruments biased by self-selection into the treatment and control? In this context, we develop a method for constructing evidence factors, and we evaluate the performance of the method in terms of design sensitivity. In the application, we consider the effectiveness of Catholic versus public high schools, constructing three evidence factors from three past strategies for studying this question. Although these three analyses use the same

data, we: (i) construct three essentially independent statistical tests that require very different assumptions, (ii) study the sensitivity of each test to the assumptions underlying that test, (iii) examine the degree to which independent tests dependent upon different assumptions concur, (iv) pool evidence across independent factors.

EO540 Room R21 STATISTICS FOR WEARABLE DEVICE DATA Chair: Jaroslaw Harezlak

E0836: **Fixed-effects inference for longitudinal functional data**
Presenter: **Ruonan Li**, North Carolina State University, United States

An inferential framework is proposed for fixed effects in longitudinal functional models and introduce tests for the correlation structures induced by sampling procedure. This provides a natural extension of standard longitudinal correlation models for scalar observations to functional observations. We compare fixed effects estimation under correctly and incorrectly specified correlation structures and provide explicit recommendations. The proposed methods are applied to the Baltimore Longitudinal Study of Aging (BLSA) to study the effects of age, sex and body mass index on the circadian rhythm of physical activity.

E0842: **Recognition of walking periods using various personal digital devices**
Presenter: **Marcin Straczkiewicz**, Harvard TH Chan School of Public Health, United States

The development of body-worn devices, such as smartphones, smartwatches, and wearable accelerometers, has remarkably deepened our understanding of how physical activity impacts human health. However, these findings are possibly just a foretaste of what data of personal digital devices may reveal as many questions on their processing and analysis remain unanswered. One such question regards activity recognition. We will introduce a novel method for quantification of walking periods from various wearable devices. We utilize the temporal dynamics of body motions measured by accelerometer. We focus on salient features of walking, namely intensity, periodicity, duration, and speed. We investigate the reflection of these phenomena at several body locations typical to wearable devices, and we create a classification scheme that allows for flexible and interpretable estimation of walking. To assess the performance of our method, we validate it over more than 300 subjects from 14 publicly available physical activity datasets. We will demonstrate that our method achieves very high classification accuracy for the recognition of walking and does not overestimate walking during other common everyday activities, regardless of sensor placement and measurement parameters.

E0965: **Estimation of free-living walking cadence from wrist-worn sensor accelerometry data**
Presenter: **Marta Karas**, Johns Hopkins Bloomberg School of Public Health, United States
Co-authors: Jacek Urbanek, Ciprian Crainiceanu, Jonas Dorn

Walking and gait parameters have become increasingly important in epidemiological and clinical studies. Indeed, 3 out of 7 submissions to the FDA for eCOA qualifications of digital endpoints are quantifying gait parameters. Recent evidence suggests that observations collected in a free-living environment are complementary to traditional, lab- and clinic-based walking measurements. Sub-second-level actigraphy data can provide a detailed description of human movement. Despite the growing need, the number of publicly available methods to derive gait from high-density accelerometry data collected by wrist-worn devices is limited. We propose an extension of ADEPT, a pattern-matching method, to segment individual walking strides in sub-second-level accelerometry data collected in a free-living environment using a wrist-worn sensor. We evaluate the method on 4-week observation data from 30 people with and 15 people without arthritis. We show that daily walking cadence is significantly associated with general mental health, social functioning, and role physical scores reported via SF-36. We provide open-source software and out-of-study sample data examples online.

E0971: **Two-sample tests for repeated measurements of histogram objects in wearable device data**
Presenter: **Haochang Shou**, University of Pennsylvania, United States
Co-authors: Jingru Zhang, Hongzhe Li

Repeated measures of wearable sensor data over multiple days have become increasingly available in biomedical research and longitudinal studies. Additionally, those data often have complex multivariate structures that are from an arbitrary non-Euclidean metric space. In particular, we will be investigating the probability densities of daily physical activity measures as densely assessed by accelerometers. Those object data are sampled from a bounded metric space and cannot be analyzed using traditional statistical methods. We propose novel non-parametric graph-based two-sample tests for the activity density object data with repeated measures. A set of test statistics are proposed to capture various possible alternatives. We derive their asymptotic null distributions under the permutation null. These tests exhibit substantial power improvements over existing methods under various alternative hypotheses and are shown to preserve the type I errors under finite samples, as shown through simulation studies. We apply the proposed tests to differentiate the distributions of daily physical activity profiles from a study population of mood disorders.

EO620 Room R22 RECENT DEVELOPMENTS IN BAYESIAN METHODOLOGY Chair: Victor Pena

E0167: **Ergodic theorems for imprecise probability kinematics**
Presenter: **Michele Caprio**, Duke University, United States
Co-authors: Sayan Mukherjee

In a standard Bayesian setting, there is often ambiguity in prior choice, as one may have not sufficient information to uniquely identify a suitable prior probability measure encapsulating initial beliefs. To overcome this, we specify a set $\mathcal{P}$ of plausible prior probability measures; as more and more data are collected, $\mathcal{P}$ is updated using Jeffrey's rule of conditioning, an alternative to Bayesian updating which proves to be more philosophically compelling in many situations. We build the sequence $(\mathcal{P}_k^*)$ of successive updates of $\mathcal{P}$ and we provide an ergodic theory to analyze its limit, for both countable and uncountable sample spaces. A result of this ergodic theory is a strong law of large numbers in the uncountable setting. We also develop a procedure for updating lower probabilities using Jeffrey's rule of conditioning.

E0244: **Dynamic matrix-variate clustering of sport activities**
Presenter: **Mattia Stival**, University of Padova-Dipartimento di Scienze Statistiche, Italy
Co-authors: Mauro Bernardi

A Bayesian matrix-variate state-space model is proposed which able to classify the trajectories of a large number (N) of P variate time series. The matrix state-space formulation allows us to consider both longitudinal and cross-sectional dependence, accounting also for missing values. Indeed, the matrix autoregressive process described by the state equation captures the time series dependence, and the use of matrix-variate normal distributions for both the measurement errors and state disturbances allows to consider cross-sectional dependence within variables (P) and between time series (N), and within states and between groups, respectively. A fully conjugate approach is adopted, and the relative Gibbs sampler is provided; to speed up the computations, Kalman recursions are performed on a vectorized and reduced form of the model. Further achievements can be derived by considering Metropolis-Hasting step to estimate in one-shot an unknown selection matrix, storing the time series cluster allocations. In the application part, we analyze the running activities of one athlete collected by his smartwatch, to say whether his performances are improving over time. In this context, data are collected as a sequence of activities, where each activity is represented by a multivariate time series, characterized by complex behavior and the presence of missing values.

E0683: **Generalized Bayesian conformal inference**
Presenter: **Federico Ferrari**, Duke University, United States

A widely touted advantage of Bayesian inference is its ability to provide a broad framework for uncertainty quantification in general settings,

including for highly complex data and models. However, in practice, Bayesian predictive intervals often have invalid out-of-sample coverage due to prior and/or model misspecification. Such a lack of 'calibration' of predictive distributions calls into question whether credible regions provide an adequate characterization of predictive uncertainty. A new generalized Bayesian framework is proposed for learning a calibrated posterior distribution that minimizes a discrepancy from an initial posterior subject to a conformal constraint encouraging validity of prediction intervals. To our knowledge, this is the first procedure in which conformal adjustments to predictive intervals, also feedback to impact parameter inferences. We propose a framework for inferring the conformal distribution that is highly flexible, easily implemented, and immediately applicable to a wide range of models. Simulation experiments and real data applications illustrate substantial gains relative to non-conformal Bayesian inference.

E0662: **Inference in response-adaptive clinical trials when the enrolled population varies over time**
Presenter: **Massimiliano Russo**, Harvard Medical School, United States
Co-authors: Steffen Ventz, Victoria Wang, Lorenzo Trippa

A common assumption of data analysis in clinical trials is that the patient population, as well as treatment effects, do not vary during the course of the study. However, when trials enrol patients over several years, this hypothesis may be violated. Ignoring variations of the outcome distributions over time, under the control and experimental treatments can lead to biased treatment effect estimates and poor control of false-positive results. We propose and compare two procedures that account for possible variations of the outcome distributions over time, to correct treatment effect estimates, and to control type I error rates the first procedure models trends of patient outcomes with splines. The second one leverages conditional inference principles, which have been introduced to analyze randomized trials when patient prognostic profiles are unbalanced across arms. These two procedures are applicable in response-adaptive clinical trials. We illustrate the consequences of trends in the outcome distributions in Bayesian response-adaptive designs and platform trials, and we investigate the proposed methods with simulations and in the analysis of a glioblastoma study.

EO227 Room R23 BAYESIAN APPLICATIONS IN BIOLOGICAL AND ENVIRONMENTAL SCIENCES Chair: Marco Ferreira

E0240: **Practical Bayesian non-parametric inference in clinical development: Single arm dose escalation studies**
Presenter: **Fabio Rigat**, Janssen R&D, United Kingdom

Dose escalation decisions in early phase clinical trials are sensitive to the functional form of the dose-response relation, with monotonicity being a key assumption. When little background information is available to inform these assumptions on biological and clinical grounds, it is unclear whether dose escalation decisions can be usefully based on regression models or whether regression-free methods should be preferred. The validity and efficiency of decisions based on a weighted average of regression estimates and observed event rates has not yet been fully assessed. These estimators are attractive in practice when the weight assigned to their regression component is estimated from the data, reflecting the models goodness of fit. Lack of widespread application of these methods for clinical trial decision making is surprising, given that they were first published as an application of Bayesian non-parametric (BNP) inference using the Dirichlet process over forty years ago. To show the potential of BNP inference for dose escalation trial design and analysis, we compare the operating characteristics of the modified continual reassessment method (mCRM) informed by either parametric or BNP estimates when the parametric model assumptions respectively do or do not hold. BNP mCRM, implemented by standard Markov chain Monte Carlo packages, is shown to offer important practical advantages when the parametric inference model is too wrong to be useful.

E0415: **Bayesian modeling of the frequency of tropical storms**
Presenter: **Joyee Ghosh**, The University of Iowa, United States

Knowing the frequency of tropical storms in advance can help in improved preparedness before a hurricane season. It is known from the existing literature that sea surface temperatures during the peak hurricane season are good predictors of tropical storm activity. However, sea surface temperatures are available after the end of the hurricane season, and thus cannot be directly used for prediction. Instead, forecasts of sea surface temperatures are available from multiple climate models. A Bayesian model averaging approach is developed that combines forecasts from multiple climate models. Based on simulation studies and North Atlantic tropical cyclone activity data, it is illustrated that this model can provide improved predictive performance compared to some existing models in the literature.

E0524: **Bayesian analysis for multi-subject time course gene expression with an application to vaccine immune response**
Presenter: **Allison Tegge**, Virginia Tech, United States
Co-authors: Marco Ferreira

A Bayesian methodology is introduced for the analysis of multi-subject time-course gene expression data. Our methodology facilitates the study of transcriptional changes through time. Specifically, we develop a fully Bayesian approach to detect differentially expressed genes that reduces the high dimensionality of time-course data by empirical orthogonal functions. The proposed model assumes distinct temporal patterns for differentially and non-differentially expressed genes, and borrows strength across genes and subjects to increase detection power. We illustrate the usefulness and flexibility of our methodology with an analysis of an RNA-seq data set from B cells to study their temporal response pattern to the human influenza vaccine.

E0724: **A novel Bayesian framework for the analysis of GWAS data**
Presenter: **Marco Ferreira**, Virginia Tech, United States
Co-authors: Jacob Williams, Tieming Ji

A Bayesian analysis is proposed for genome-wide association studies (GWAS) that selects significant single nucleotide polymorphisms (SNP) in two stages. The first stage fits as many linear mixed models as the number of SNPs, with each model containing one SNP as well as random effects to take into account kinship correlation. The result of the first stage is a set of candidate significant genes. The second stage performs a stochastic search through model space with a genetic algorithm, where each model is a linear mixed model with principal components for population structure as well as kinship random effects and may include multiple SNPs from the candidate set obtained from the first stage. The result of the second stage is a list of models with their respective posterior probabilities. We illustrate our proposed Bayesian GWAS framework with an analysis of publicly available experimental data on root architecture remodeling of a model plant in response to salt stress.

EO377 Room R24 SAMPLING AND CORESETS FOR LARGE-SCALE DATA Chair: Jingshen Wang

E0318: **Bootstrapping max statistics in high dimensions**
Presenter: **Miles Lopes**, UC Davis, United States
Co-authors: Hans-Georg Mueller

In recent years, bootstrap methods have drawn attention for their ability to approximate the laws of max statistics in high-dimensional problems. A leading example of such a statistic is the coordinate-wise maximum of a sample average of n random vectors in R^p. Existing results for this statistic show that the bootstrap can work when $n << p$, and rates of approximation (in Kolmogorov distance) have been obtained with only logarithmic dependence in p. Nevertheless, one of the challenging aspects of this setting is that established rates tend to scale slower than $n^{-1/2}$ as a function of n. The main purpose is to demonstrate that improvement in rate is possible when extra model structure is available. Specifically, we show that if the coordinate-wise variances of the observations exhibit decay, then a nearly $n^{1/2}$ rate can be achieved, independent of p. Furthermore, a surprising

aspect of this dimension-free rate is that it holds even when the decay is very weak. Lastly, we provide examples showing how these ideas can be applied to inference problems dealing with functional and multinomial data.

E0653: **Maximum sampled likelihood estimation for informative subsampling**
Presenter: **HaiYing Wang**, University of Connecticut, United States
Co-authors: Jae Kwang Kim

Subsampling is an effective approach to extract useful information from massive data sets when computing resources are limited. Existing investigations focus on developing better sampling procedures and deriving probabilities with higher estimation efficiency. After a subsample is taken from the full data, most available methods use an inverse probability weighted target function to define the estimator. This type of weighted estimator reduces the contributions of more informative data points, and thus it does not fully utilize information in the selected subsample. The focus is on parameter estimation with a selected subsample. We propose to use the maximum sampled likelihood estimator (MSLE) based on the sampled data. We established the asymptotic normality of the MSLE, and prove that its variance-covariance matrix reaches the lower bound of asymptotically unbiased estimators. Specifically, the MSLE has a higher estimation efficiency than the weighted estimator. We further discuss the asymptotic results with the L-optimal subsampling probabilities. We illustrate the estimation procedure with generalized linear models. Numerical experiments are provided to evaluate the practical performance of the proposed method.

E0804: **Sparse variational inference**
Presenter: **Trevor Campbell**, University of British Columbia, Canada

The purpose is to cover recent work on Bayesian coresets (core of a dataset), a methodology for statistical inference via data compression. Coresets achieve compression by forming a small weighted subset of data that replaces the full dataset during inference, leading to significant computational gains with provably minimal loss in inferential quality. In particular, the talk will present methods for Bayesian coreset construction, from previously-developed subsampling, greedy, and sparse linear regression-based techniques to a novel algorithm based on sparse variational inference (VI). In contrast to past algorithms, sparse VI is fully automated, requiring only the dataset and probabilistic model specification as inputs. Empirical results will be shown which illustrate that despite requiring much less user input than past methods, sparse VI coreset construction provides state-of-the-art data summarization for Bayesian inference.

E1095: **Discussant**
Presenter: **Jingshen Wang**, University of Michigan, United States

Based on the session speakers' recent work, relevant progress in the field of sampling for large-scale data will be discussed.

EO317 Room R25 STATISTICAL MODELS: RECENT DEVELOPMENTS I Chair: Anna Panorska

E0883: **Omnibus test for normality based on the Edgeworth expansion**
Presenter: **Agnieszka Wylomanska**, Wroclaw University of Science and Technology, Poland

Statistical inference in the form of hypothesis tests and confidence intervals often assumes that the underlying distribution is normal. Similarly, many signal processing techniques rely on the assumption that a stationary time series is normal. As a result, a number of tests have been proposed in the literature for detecting departures from normality. We develop a novel approach to the problem of testing normality by constructing a statistical test based on the Edgeworth expansion, which approximates a probability distribution in terms of its cumulants. By modifying one term of the expansion, we define a test statistic which includes information on the first four moments. We perform a comparison of the proposed test with existing tests for normality by analyzing different platykurtic and leptokurtic distributions including generalized Gaussian, mixed Gaussian, alpha-stable and Students-t distributions. We show for some considered sample sizes that the proposed test is superior in terms of power for the platykurtic distributions whereas for the leptokurtic ones it is close to the best tests like those of DAgostino-Pearson, Jarque-Bera and Shapiro-Wilk.

E0938: **The size effect in the stock market**
Presenter: **Andrey Sarantsev**, University of Nevada in Reno, United States

Returns of equally-weighted size deciles of the USA stock market from the Center for Research in Securities Prices database 1926-2020 are compared. We study dependence of these returns upon market weights and the benchmark = the top decile. We construct a simple time series model and study its properties.

E1011: **Empirical anomaly measure for finite-variance processes**
Presenter: **Katarzyna Maraj**, Wroclaw University of Science and Technology, Poland

Anomalous diffusion phenomena are observed in many areas of interest. They manifest themselves in deviations from the laws of Brownian motion (BM), e.g. in the non-linear growth (mostly power-law) in time of the ensemble average mean squared displacement (MSD). When we analyze the real-life data in the context of anomalous diffusion, the primary problem is the proper identification of the type of anomaly. We introduce the new statistic, called the empirical anomaly measure (EAM) that can be useful for this purpose. This statistic is the sum of the off-diagonal elements of the sample autocovariance matrix for the increments process. On the other hand, it can be represented as the convolution of the empirical autocovariance function (ACVF) with time lags. The EAM statistic measures interdependence between the ensemble-averaged MSD of the given process from the ensemble-averaged MSD of the classical BM. We prove the main probabilistic characteristics of the EAM statistic and construct the formal test for the recognition of the anomaly type. The advantage of the EAM is the fact that it can be applied to any data trajectories without the model specification. The only assumption is the stationarity of the increments process. The complementary summary of the presentation constitutes of Monte Carlo simulations illustrating the effectiveness of the proposed test and properties of EAM for selected processes.

E1024: **Use of compound Poisson-gamma distribution as a model for precipitation: A case study from the south of Sweden**
Presenter: **Anastassia Baxevani**, University of Cyprus, Cyprus
Co-authors: Christos Andreou

Many statistical models exist for modelling precipitation. The main challenge when modelling precipitation is that one needs to model both the presence of exact zeros that correspond to dry days and the amounts of precipitation on the wet days. We suggest using the compound Poisson gamma distribution, which belongs to the family of distributions known as the Tweedie family, and has the ability to produce exact zeros and marginals that follow the gamma distribution. We discuss different methods for fitting the *compound Poisson gamma* distribution to daily precipitation data from southern Sweden. Then, we generate realistic artificial precipitation sequences to estimate different precipitation characteristics.

CI023 Room R02 RECENT ADVANCES IN TIME SERIES ECONOMETRICS Chair: Peter Pedroni

C0228: **Inference in time series models using smoothed-clustered standard errors**
Presenter: **Timothy Vogelsang**, Michigan State University, United States
Co-authors: Seunghwa Rho

A long run variance estimator is proposed for conducting inference in time series regression models that combines the nonparametric approach with a cluster approach. The basic idea is to divide the time periods into non-overlapping clusters. The long run variance estimator is constructed by first aggregating within clusters and then kernel smoothing across clusters or applying the nonparametric series method to the clusters with Type

II discrete cosine transform. We develop asymptotic theory for test statistics based on these "smoothed-clustered" long run variance estimators. We derive asymptotic results holding the number of clusters fixed and also treating the number of clusters as increasing with the sample size. For kernel smoothing, these two asymptotic limits are different whereas for the cosine series approach, the limits are the same. When clustering before kernel smoothing, the "fixed-number-of-clusters" asymptotic approximation works well whether the number of clusters is small or large. Finite sample simulations suggest that the naive i.i.d. bootstrap mimics the fixed-number-of-clusters critical values. The simulations suggest that clustering before kernel smoothing can reduce over-rejections caused by strong serial correlation with a cost of power. When clustering is natural, it can reduce over-rejection problems and achieve small gains in power for the kernel approach. In contrast, the cosine series approach does not benefit from clustering.

C0288: **Inference on the dimension of the nonstationary subspace in functional time series**
Presenter: **Morten Nielsen**, Queen's University, Canada
Co-authors: Won-Ki Seo, Dakyung Seong

A statistical procedure is proposed to determine the dimension of the nonstationary subspace of cointegrated functional time series taking values in the Hilbert space of square-integrable functions defined on a compact interval. The procedure is based on sequential application of a proposed test for the dimension of the nonstationary subspace. To avoid estimation of the long-run covariance operator, our test is based on a variance ratio-type statistic. We derive the asymptotic null distribution and prove consistency of the test. Monte Carlo simulations show good performance of our test and provide evidence that it outperforms the existing testing procedure. We apply our methodology to three empirical examples: age-specific US employment rates, Australian temperature curves, and Ontario electricity demand.

C0172: **Robust estimation of long run functions of unknown form in panel time series**
Presenter: **Peter Pedroni**, Williams College, United States
Co-authors: Stephan Smeekes

A new technique is investigated for estimating specialized long run functional relationships. The technique uses polynomial approximations to estimate functions of unknown form. It exploits the structure of unit root panels by decomposing the estimating equations into a set of static linear time series regressions for each unit followed by a set of cross-sectional polynomial regressions for each historical period of observation. We establish asymptotic normality and fast rates of convergence that allow for considerable robustness with respect to temporal and cross-sectional dependency, including common unit root factors. We also investigate the attractive finite sample properties of the technique by Monte Carlo simulation. We offer two illustrations of empirical application, one from the growth literature and one pertaining to the environmental Kuznets curve.

CO149 Room R04 SENTOMETRICS Chair: Keven Bluteau

C0448: **FiGAS: Fine-Grained Aspect-based Sentiment analysis on economic and financial lexicon**
Presenter: **Sergio Consoli**, National Research Council of Italy (CNR), Italy
Co-authors: Luca Barbaglia, Sebastiano Manzan

The extraction of sentiment from news, social media and blogs for the prediction of economic and financial variables has attracted great attention in recent years. Despite many successful applications of sentiment analysis (SA) in these domains, the range of semantic techniques employed is still limited and mostly focus on the detection of sentiment at a coarse-grained level, i.e. whether the sentiment expressed by the entire sentence text is either positive or negative. However, coarse-grained methods might not be precise enough in evaluating the sentiment polarity of a specific topic of interest contained in a sentence. For this reason we propose FiGAS, a Fine-Grained Aspect-based Sentiment analysis approach that is able to identify the sentiment associated to specific topics of interest within a text, by assigning real-valued sentiment polarity scores to those topics. The approach is completely unsupervised and customized to the economic and financial domains, being built upon a large specialised lexicon in these areas. We provide a statistical comparison of the performance of FiGAS against other popular lexicon-based SA approaches on a humanly annotated data set in the economic and financial domain. FiGAS outperforms the other methodologies and shows to be a promising alternative to extract sentiment from news.

C0689: **Joint sent/topic modelling: The non-exchangeability of topics trap and improved convergence diagnostics**
Presenter: **Olivier Delmarcelle**, Ghent University, Belgium
Co-authors: Kris Boudt, David Ardia

The joint sentiment/topic model (JST) aims at opinion mining textual data by estimating in an unsupervised way jointly the topics and sentiment in a text, while allowing the sentiment classification to be conditional on the topic. We warn users of JST models against convergence issues due to the trap of non-exchangeability of sentiment/topics leading to a large number of modes polluting the MCMC inference. This pitfall becomes evident when expressing the JST model under the novel view of the Dirichlet-Tree distribution. We propose a coherence metric designed to evaluate the quality of the inferred models. Experiments on synthetic data conclude that this metric is better at estimating model quality than likelihood. Finally, we provide general guidelines on the usage of JST-class models and the tuning of their hyper-parameters.

C0820: **Nowcasting GDP growth using media news articles**
Presenter: **Arno De Block**, Vrije Universiteit Brussel, Belgium
Co-authors: Andres Algaba, Geert Langenus, Peter Reusens

GDP growth is the key indicator for measuring the state of the economy. It is closely monitored by policymakers, firms, and investors to optimize their economic decision-making. But while decisions have to be made in real-time, GDP growth is only observed at a quarterly frequency and with a substantial publication lag. To monitor the state of the economy in real-time, high-frequency macroeconomic and financial variables are used to nowcast GDP growth. We propose to augment this real-time dataset with information embedded in media news articles. The regression-based nowcasting approaches allow us to construct media-based variables that are optimal for the application of nowcasting GDP growth. Moreover, by enforcing sparsity, we manage only to select the most informative macroeconomic, financial, and media-based variables leading to superior nowcasting performance.

C1121: **Climate change concerns and the performance of green versus brown stocks**
Presenter: **Keven Bluteau**, HEC Montreal, Canada
Co-authors: David Ardia, Kris Boudt, Koen Inghelbrecht

The aim is to empirically test the prediction that green firms can outperform brown firms when climate change concerns strengthen unexpectedly for S&P 500 companies over the period of January 2010 - June 2018. To capture unexpected increases in climate change concerns, we construct a Media Climate Change Concern index using climate change-related news published by major U.S. newspapers. We find a negative relationship between the firms' exposure to the Media Climate Change Concerns index and the level of the firm's greenhouse gas emission per unit of revenue. This result implies that when concerns about climate change rise unexpectedly, green firms' stock price increases, while brown firms' stock price decreases. Further, using topic modeling, we analyze which type of climate change news drives this relationship. We identify five themes that affect green vs brown stock returns. Some of those themes can be related to change in investors' expectations about the future cash-flow of green vs brown firms, while others cannot. This result implies that the relationship between concern and green vs brown stock returns arises from both investors updating their expectations about the future cash-flows of green and brown firms and changes in investors' sustainability taste.

CO193 Room R06 APPLIED ECONOMETRICS Chair: Michael Owyang

C0210: **A composite index for evaluating the stance of monetary policy**
Presenter: **Laura Jackson Young**, Bentley University, United States
Co-authors: Michael Owyang, David Wheelock

The Federal Reserve's response to the financial crisis and the Great Recession involved a multi-instrument approach to monetary policy. In addition to lowering the federal funds target rate to nearly zero, the Fed also increased the size and changed the composition of the balance sheet, employed forward guidance, and targeted longer-term yields. Under the assumption that the federal funds target is the only (or at least primary) policy instrument, one can assess the effects of monetary policy by examining shocks to the fed funds rate. The use of multiple instruments has complicated models of monetary policy as one cannot simply consider the effects of a shock to a single variable. We develop an empirical model in which monetary policy is summarized by a single latent series. Our model uses a factor structure that can combine a variety of instruments. We adopt a version of the factor-augmented VAR (FAVAR) with time-varying loadings. Our assumption is that monetary policy, when properly measured, affects the economy linearly. However, the stance of policy is an agglomeration of a number of different measures with time-varying weights.

C0236: **Reconsidering the Feds forecasting advantage**
Presenter: **Amy Guisinger**, Lafayette College, United States
Co-authors: Michael Owyang, Michael McCracken

Previous studies have found that the Federal Reserve's (Greenbook) forecasts of inflation are superior to that of professional forecasters. More recent papers, however, suggest that this advantage has been dissipating over time. We investigate the origin of the Fed's forecasting advantage, focusing on an explanation that was previously dismissed, that the Fed's forecasts are conditional on the path of policy, about which the Fed may have more information. To do this, we examine whether the Federal Reserve's advantage remains after controlling for information about future monetary policy. We find that the Feds forecasts no longer encompasses the private sectors once accounting for the future path of policy, regardless of the subsample used for estimation. We then consider whether the Fed's advantage remains when the market is operating under the same (or similar) beliefs as the Fed. We identify dates when the markets expectation of future policy coincided with the Greenbooks conditioning path and find the Fed seems to possess no informational advantage during periods where the markets expectations of future policy are synched to the Feds.

C0266: **Capital flows in risky times: Risk-on/risk-off and emerging market tail risk**
Presenter: **Karlye Dilts Stedman**, Federal Reserve Bank of Kansas City, United States
Co-authors: Anusha Chari, Christian Lundblad

The implications of risk-on/risk-off shocks for emerging market capital flows and returns are characterized. We document that these shocks have important implications, not only for the median of emerging markets flows and returns, but also for the left tail. Further, while there are some differences in the effects across bond vs. equity markets and flows vs. asset returns, the effects associated with the worst realizations are generally larger than that on the median realization. We apply our methodology to the COVID-19 shock to examine the pattern of flow and return realizations: the sizable risk-off nature of this shock engenders reactions that reside deep in the left tail of most relevant emerging market quantities.

C0513: **On the reliability of Central Banks' fancharts: Calibration of density path forecasts**
Presenter: **Giulia Mantoan**, University of Warwick, United Kingdom
Co-authors: Ana Galvao, James Mitchell

Central Banks regularly publish fan charts of macroeconomic variables, communicating forecasts for several horizons. However, fan charts contain three types of information: point forecasts (the path), likelihood of the path (bands around it), and variable's dynamics across horizons. Existing absolute evaluation approaches neglect the latter. Practitioners indeed evaluate the calibration of fan charts testing the forecast accuracy horizon by horizon, not considering any time-dependency among them. Fan charts are described as density path forecasts, the impact of time-dependence in the evaluation is discussed, and calibration tests are proposed to assess whether or not Central Banks publishes reliable forecasts.

CO309 Room R07 SELF-FULFILLING PROPHECIES AND MACROECONOMIC BEHAVIOR Chair: Marco Maria Sorge

C0483: **On model predictions for forward guidance puzzle**
Presenter: **Nigel McClung**, Bank of Finland, Finland

Sufficient conditions are provided for when a rational expectations structural model predicts bounded responses of endogenous variables to forward guidance announcements. The conditions coincide with a special case of the well-known (E)xpectation-stability conditions that govern when agents can learn a Rational Expectations Equilibrium. The conditions are distinct from the determinacy conditions and are applicable in a wide variety of models, including models with sunspots or Markov-switching. We show how the conditions can diagnose features of a model that contribute to the Forward Guidance Puzzle and reveal how to construct well-behaved forward guidance predictions in standard medium-scale New Keynesian environments.

C0267: **On rational fat tails**
Presenter: **Chetan Dave**, University of Alberta, Canada
Co-authors: Marco Maria Sorge

The purpose is examine the role of sunspot shocks in generating fat-tailed behavior of endogenous variables in equilibrium business cycle models without departing from the Gaussian rational expectations (RE) benchmark. We formally establish that any RE model exhibiting indeterminacy admits a linear recursive equilibrium representation as a function of regularly varying multiplicative noise (LRMN). This, in turn, allows small shocks to fuel large deviations, thereby imparting non-Gaussian features to equilibrium patterns in standard Gaussian environments. Numerical simulations and an estimation exercise highlight the ability of LRMN representations to replicate statistical regularities with respect to fat-tailed distributions and high-probability extreme outcomes.

C0389: **On irrationality or risk aversion: A sticky expectations model for households' unemployment predictions**
Presenter: **Luca Gerotto**, Universita Cattolica del Sacro Cuore, Italy
Co-authors: Paolo Pellizzari

A novel theoretical framework is developed that combines a staggered update of information with risk-aversion of the individuals. The combination of these two features leads a risk-averse individual, conscious of having outdated information, to rationally under(over)-estimate a variable in such a way to hedge against a costly over(under)-estimation. Model parameters are calibrated on empirical data: specifically, the focus is on unemployment expectations of Italian households, from the Consumer Confidence Survey run by *ISTAT* on Italian households. The findings suggest that the heterogeneity in the frequency of update of information is the main driver of the different aggregate bias for different demographic groups. This frequency is increasing in the education level, is higher for men than for women and is higher for working people (in particular the self-employed) than for not-working individuals. It is also lower in the Southern regions of Italy and is higher for young agents than for retired people. This result is important for policy-making since it allows to identify the groups which are more in need of easy access to information as well as financial (and economic) education.

C0439: **Under the same (Chole)sky: DNK models with 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 questioned the reliability of recursive identification schemes based on short-run exclusion restrictions. We track down the role of informational constraints embodying classical Cholesky-timing restrictions in shaping equilibrium reduced forms of otherwise standard DSGE models. These are found to engender (i) invertible (thereby fundamental) equilibrium representations for the observables under plausible parameterizations, and (ii) theoretical impulse response functions (IRFs) that may differ sharply, over nearly all time horizons, from those implied by unrestricted model counterparts, conditional on the model's internal propagation mechanism. As a result, the Cholesky identification in a truly Cholesky (recursive) world need not distort estimated monetary impulse responses, allowing short-run restrictions to serve as a useful identification tool.

CO544 Room R08 TOPICS IN MACROECONOMETRICS Chair: Tatevik Sekhposyan

C1142: **The efficiency of the Eurosystem/ECB staff inflation projections: A state-dependent analysis**
Presenter: **Eleonora Granziera**, Norges Bank, Norway
Co-authors: Maritta Paloviita, Pirkka Jalasjoki

The Eurosystem/ECB staff inflation projections are examined to assess whether new information is efficiently incorporated in the forecasting process. We use a confidential dataset of the ECB staff macroeconomic quarterly projections and determine if there was an efficient use of information by testing whether the forecast error is predictable, given the state of the economy. Specifically, we investigate whether systematic errors are related to the behaviour of inflation when the forecasts are made and distinguish whether inflation is above or below target at the time the forecasts are made. Our analysis suggests that the forecast errors are unbiased on average; however, there is evidence of state dependence. In particular, we find that the ECB tends to overpredict (underpredict) inflation when inflation is below (above) target. This result holds even after accounting for errors in the external assumptions.

C1154: **Measuring aggregate and sectoral uncertainty**
Presenter: **Luis Uzeda**, Bank of Canada, Canada
Co-authors: Efrem Castelnuovo, Kerem Tuzcuoglu

An empirical framework is proposed for the estimation of aggregate and sectoral uncertainty that is suitable for data-rich environments. Building upon previous works on multilevel factor and common stochastic volatility models, we jointly decompose the conditional mean and variance of economic time series into an aggregate and a sectoral component. Uncertainty – aggregate and sectoral – is captured by common factors driving the conditional variance of economic variables at different levels of aggregation in a large system. We apply this methodology to two large datasets for the U.S. economy. The results indicate that, while aggregate uncertainty dominates during sharp recessions, since the mid-1980s sectoral uncertainty has become more salient than its aggregate counterpart. An early assessment of the ongoing Covid-19 pandemic through the lens of our framework indicates that the current crisis led to an unprecedented spike of aggregate uncertainty, more than double compared with the Great Recession. Also, the durable and non-durable manufacturing sectors witnessed substantial sector-specific uncertainty at the onset of the pandemic.

C1153: **Central bank density forecasts and asset prices: Do revisions to higher-order moments matter?**
Presenter: **Tatevik Sekhposyan**, Texas A and M University, United States
Co-authors: Ryan Rholes

The Bank of England's density forecasts and its revisions are considered to quantify the effects of information flow on the financial markets. This fits into the broader literature on the effects of macroeconomic news on financial markets. It is, however, more particularly interesting from a monetary policy perspective. Forward guidance, i.e. communication about the state of the economy and likely future course of monetary policy, has become an increasingly important part of the central banks' toolkit around the world. These communications are often in the form of published forecasts. Point forecasts and their revisions have been shown to move the financial markets. The effects of the higher moments, however, have not been investigated thoroughly, and this is primarily due to data limitations. Bank of England, on the other hand, has been publishing information on its density forecasts since the late 1990s, making it useful for our analysis. Using daily information on the financial markets, we find that the updates of higher moments are more important in moving the financial markets than the revisions to the first central moment of the density forecasts, making them relevant for the effectiveness of the monetary policy.

C1146: **Climate change and social inequality**
Presenter: **Tatjana Dahlhaus**, Bank of Canada, Canada

The effects of climate change on inequality are assessed. First, we model climate change as shifts in the whole distribution of weather variables (e.g., temperatures). In doing so, we calculate time series of, for example, mean, variance, max-min, interquantile range, skewness, kurtosis, Value at Risk (VaR), Quantiles, etc. Next, we proceed by describing the time series pattern in these series, such as the presence of trends. Finally, the paper builds an empirical structural model to model economic variables and climate change endogenously. Specifically, we introduce a functional VAR and define climate change as a shock that permanently shifts the time-varying distribution of weather variables. Potential economic variables of interest include GDP, but also social inequality. We focus on inequality and study the distributional effects of climate change on income and wealth within the US. Inequality has been a persistent issue in the climate change discussion with a focus on inequality across countries. Nevertheless, within-country inequality has not received much attention, and we aim to provide first insights into the effects of climate change on income inequality within a country.

CG613 Room R03 CONTRIBUTIONS IN PORTFOLIO ANALYSIS AND ASSET PRICING Chair: Mohammad Jahan-Parvar

C0411: **A basket half full: Sparse portfolios**
Presenter: **Ekaterina Seregina**, University of California, Riverside, United States

The existing approaches to sparse wealth allocations (1) are suboptimal due to the bias induced by ℓ_1-penalty; (2) require the number of assets to be less than the sample size; (3) do not model factor structure of stock returns in high dimensions. We address these shortcomings and develop a novel strategy which produces unbiased and consistent sparse allocations. We demonstrate that: (1) failing to correct for the bias leads to low out-of-sample portfolio return; (2) only sparse portfolios achieved positive cumulative return during several economic downturns, including the dot-com bubble of 2000, the financial crisis of 2007-09, and COVID-19 outbreak.

C0822: **Copula-based multiobjective portfolio optimization**
Presenter: **Maziar Sahamkhadam**, Linnaeus University, Sweden

Utilizing the multicriteria decision making (MCDM) and vine copulas, a copula-based multiobjective portfolio (MOP) optimization is developed. Using the copula-based MOP optimization model, we evaluate the impact of objective functions on portfolio performance. Furthermore, we compare the copula-based MOP with those obtained from two sophisticated predictive models, including the multivariate GARCH and the factor stochastic volatility. Applying the MOP optimization to a sample of S&P 100 equities, an in-sample analysis of the Pareto sets reveals that the risk models generally perform better than a historical-simulation approach in generating optimal efficient sets when there are higher preferences on return and tail risk. Based on an out-of-sample analysis, the predictive models generate multicriteria portfolios that achieve statistically significant improvements concerning return and tail risk during a market downturn such as the global financial crisis (GFC) and the COVID-19 market crash.

Overall, there is evidence that the copula-based multicriteria portfolios perform better than those from the other predictive models in terms of the downside risk. With a view to the portfolio attributes, the dividend yield and beta coefficient show significant negative impacts on portfolio tail risk measures.

C0319: **Dynamic currency hedging using non-Gaussian returns model**

Presenter: **Urban Ulrych**, University of Zurich and Swiss Finance Institute, Switzerland
Co-authors: Pawel Polak

Managing a portfolio with foreign currency exposure is a critical aspect of international asset allocation. A new foreign currency hedging strategy for international investors is motivated and studied. In the theoretical part of the work we start with the model-free optimal currency exposures and assume a very flexible non-Gaussian returns model for currency and portfolio returns. In the context of our model, each element of the vector return at time t is endowed with a common univariate shock, interpretable as a common market factor. We show that this mixing random variable plays the role of ambiguity (uncertainty about the return distribution), whereby its magnitude is expressed through the size of the market factor's conditional variance. Building on the derived theoretical model we propose a semi-parametric extended filtered historical simulation approach to model the future distribution of asset and currency returns. Based on this, we introduce an algorithm for dynamic currency hedging that can be used to numerically optimize any coherent risk measure, such as Expected Shortfall (ES). The out-of-sample back-test results show that the optimal ES hedging strategy outperforms the benchmarks of constant hedging as well as equivalent approaches based on GARCH modelling net of transaction costs.

C0691: **Bayesian quantile-based portfolio selection**

Presenter: **Vilhelm Niklasson**, Stockholm University, Sweden
Co-authors: Taras Bodnar, Erik Thorsen, Mathias Lindholm

The optimal portfolio allocation problem is studied from a Bayesian perspective using value at risk (VaR) and conditional value at risk (CVaR) as risk measures. By applying the posterior predictive distribution for the future portfolio return, we derive relevant quantiles needed in the computations of VaR and CVaR, and express the optimal portfolio weights in terms of observed data only. This is in contrast to the conventional method where the optimal solution is based on unobserved quantities which are estimated, leading to suboptimality. We also obtain the expressions for the weights of the global minimum VaR and CVaR portfolios, and specify conditions for their existence. Moreover, analytical expressions for the mean-VaR and mean-CVaR efficient frontiers are presented and the extension of theoretical results to general coherent risk measures is provided. One of the main advantages of the suggested Bayesian approach is that the theoretical results are derived in the finite-sample case and thus they are exact and can be applied to large-dimensional portfolios. By using simulation and real market data, we compare the new Bayesian approach to the conventional method by studying the performance and existence of the global minimum VaR portfolio and by analysing the estimated efficient frontiers. It is concluded that the Bayesian approach outperforms the conventional one, in particular at predicting the out-of-sample VaR.

Monday 21.12.2020 08:45 - 10:00 Parallel Session M – CFE-CMStatistics

EO700 Room R04 COPULA MODELS IN ECONOMETRICS Chair: Ralf Wilke

E0166: **A copula duration model for multiple-states-multiple-spells**
Presenter: **Shuolin Shi**, Copenhagen Business School, Denmark
Co-authors: Ralf Wilke, Ming Sum Simon Lo

Copula graphic estimator for competing risks duration model utilizes the Archimedean copula to model the dependence structure of risk-specific durations. The nested Archimedean copula function permits for different dependence structures between risks, spells, and states, and it does not put any restriction on the cumulative incidence function. The dependence structure between risks is not identified and can only be assumed, yet the dependence structures between spells and states can be identified and estimated. So, the copula graphic estimator for multiple-states-multiple-spells competing risks duration model is flexible and can serve as a tool to assess the relevance of assumptions on the dependence structure of risk-specific durations. The estimation procedure is worked out and tested with both simulations and empirical Danish administrative data. The estimator is still consistent if we ignore the dependence structure, yet a loss of efficiency is expected.

E0196: **Dynamic relationship between stock market and bond market: A GAS-MIDAS copula approach**
Presenter: **Hoang Nguyen**, Orebro University, Sweden
Co-authors: Farrukh Javed

There is evidence that macroeconomic variables influence the relationship among financial variables, however they are sampled at different frequencies. A generalized autoregressive score mixed frequency data sampling (GAS-MIDAS) copula approach is proposed to analyze the dynamic relationship between the Stock market and the Bond market. A GAS-MIDAS copula decomposes their dependence into a short-run and a long-run correlation. While the long term effect is updated at a lower frequency using MIDAS, the short term effect is taken into account using a GAS approach. The model helps to improve the in-sample and the out-of-sample forecast.

E0392: **Competing risks regression with dependent multiple spells: Monte Carlo evidence and an application to maternity leave**
Presenter: **Ralf Wilke**, Copenhagen Business School, Denmark
Co-authors: Ming Sum Simon Lo, Shuolin Shi, Caecilia Lipowski

Copulas are a convenient tool in statistics to model dependencies. A dependent competing risks model with dependent multiple spells is considered. The aim is to study the practical model performance depending on the choice of hazard model and copula. A simulation study looks at the relevance of the assumed parametric or semiparametric models for hazard functions, copula and whether full or partial maximum likelihood approach is chosen. The results show that the researcher must be careful which hazard is being specified as similar functional form assumptions for the subdistribution and cause-specific hazard will lead to different estimated cumulative incidences. Model selection tests for the choice of hazard model and copula are found to provide some guidance for setting up the model. The nice practical properties and flexibility of the copula model are demonstrated with an application to a large set of maternity leave of mothers after they have given birth to their first to up to the third child.

EO145 Room R05 ADVANCES IN STATISTICAL MODELLING Chair: Efoevi Angelo Koudou

E0169: **Relative variation indexes for multivariate positive continuous distributions**
Presenter: **Aboubacar Y Toure**, Universite Bourgogne Franche-Comte, France
Co-authors: Celestin C Kokonendji, Amadou Sawadogo

Some new indexes are introduced to measure the departure of any multivariate continuous distribution on the nonnegative orthant of the corresponding space from a given reference distribution. The reference distribution may be an uncorrelated exponential model. The proposed multivariate variation indexes that are a continuous analogue to the relative Fisher dispersion indexes of multivariate count models are also scalar quantities, defined as ratios of two quadratic forms of the mean vector to the covariance matrix. They can be used to discriminate between continuous positive distributions. Generalized and multiple marginal variation indexes with and without correlation structure, respectively, and their relative extensions are discussed. The asymptotic behaviors and other properties are studied. Illustrative examples, as well as numerical applications, are analyzed under several scenarios, leading to appropriate choices of multivariate models. Some concluding remarks and possible extensions are made.

E0402: **TailCoR: A new measure of dependence**
Presenter: **Sladana Babic**, Ghent University, Belgium
Co-authors: Christophe Ley, Lorenzo Ricci, David Veredas

Economic and financial crises are characterized by unusually large events. These tail events co-move because of linear and/or nonlinear dependencies. We introduce TailCoR, a metric that combines (and disentangles) these linear and non-linear dependencies. TailCoR between two variables is based on projecting them into a line and computing a tail Inter Quantile Range of the projection. TailCoR is therefore simple to compute. Moreover, no optimizations are needed, it is dimension-free, and it performs well in small samples.

E0962: **About the Stein equation related to the generalized inverse Gaussian and Kummer distributions**
Presenter: **Efoevi Angelo Koudou**, IECL CNRS /Universite de Lorraine, France
Co-authors: Essomanda Konzou, Kossi Gneyou

By using a general approach existing in the literature for distributions satisfying a certain differential equation, a new bound is established for the solution of the Stein equation related to the generalized inverse Gaussian (resp. the Kummer) distribution. This bound is optimal for Lipschitz test functions. It has an explicit expression as a function of the parameters of the distribution in terms of the modified Bessel function of the third kind (resp. the confluent hypergeometric function of the second kind).

EO466 Room R11 CHALLENGES IN FUNCTIONAL DATA ANALYSIS AND VARYING COEFFICIENT MODELS Chair: Yuko Araki

E0482: **Penalized kernel quantile regression for varying coefficient models**
Presenter: **Eun Ryung Lee**, Sungkyunkwan University, Korea, South

Kernel smoothing has been a prevalent and successful nonparametric estimation method in the literature. Nevertheless, a penalization technique in kernel smoothing has been poorly understood due to its intrinsic technical and computational issues. We develop a novel penalized and computational method working with local linear quantile regression in varying coefficient models. We show that the proposed method consistently identifies the partially linear structure of the varying coefficient model even when the number of covariates is allowed to increase with the sample size. We develop an efficient algorithm using the alternating direction method of multipliers with a computational convergence guarantee. Such development requires novel technical and computational analyses, especially when the dimension diverges to infinity. The proposed method outperforms the existing penalized kernel algorithms, and it can be easily extended to other problems in penalized kernel smoothing. The various simulations and real data analyses conducted illustrate the effectiveness and numerically verify the proposed method.

E0541: **Variable selection in varying-coefficient functional linear models**
Presenter: **Hidetoshi Matsui**, Shiga University, Japan

Varying-coefficient functional linear models consider the relationship between a scalar response and functional predictors, where the coefficient

functions depend on an exogeneous variable. It then accounts for the relation of the predictors and the response varying with the exogeneous variable. We consider the problem of variable selection in the varying-coefficient functional linear model with multiple functional predictors. To solve this problem, we apply the group lasso-type regularization that induces sparsity. The proposed method is applied to the analysis of agricultural data. In particular, we select environmental factors that relate to the crop yield of multi-stage tomatoes.

E1021: **Functional survival analysis with several classes of trajectories**
Presenter: **Yuko Araki**, Shizuoka University, Japan

Survival analysis with a few classes of functional covariates is considered. There are several approaches to derive the class of latent trajectories from longitudinal observation, such as latent class growth models. We first identify the class of individual trajectory by simple functional clustering method, and then we use this information as covariates in survival analysis. In a simulation study, we compare the proposed method to other latent trajectory approaches. For illustration, we use Body Mass Index data observed for a few decades in Japan.

EO754 Room R12 SOME RECENT STATISTICAL DEVELOPMENTS IN CLUSTERING Chair: Audrey Poterie

E0562: **Statistical analysis of a hierarchical clustering algorithm with outliers**
Presenter: **Audrey Poterie**, Universite Bretagne Sud, France
Co-authors: Laurent Rouviere, Nicolas Klutchnikoff

In unsupervised learning, the single linkage is a hierarchical clustering method which consists in recursively merging the two closest clusters in term of minimal distance. Even if this procedure has many interesting properties, it is well known that, due to the chaining problem, the procedure usually fails to identify clusters in the presence of outliers (observations that do not belong to any clusters). We propose a new version of this algorithm and we study its mathematical performances. In particular, we provide an oracle inequality which ensures that the proposed procedure is efficient under mild assumptions on the size of the clusters. The performances of our approach are also assessed through a simulation study involving various synthetic data sets and a comparison with some classical clustering algorithms is also presented.

E0610: **Multivariate functional data clustering using unsupervised binary trees**
Presenter: **Steven Golovkine**, CREST, France
Co-authors: Nicolas Klutchnikoff, Valentin Patilea

With the recent development of sensing devices, more and more data are recorded in both dimensions of time and space. These measures lead to large amounts of data that are often referred to as multivariate functional data. A simple clustering procedure is proposed for such multivariate functional data. Considering a multivariate functional principal components analysis as a dimension reduction vehicle, a binary tree is grown using a parametric mixture model defined on the projection of the trajectories onto the principal components. The mixture model is fitted by an EM algorithm. Then, a joining step is introduced to eventually merge the similar nodes of the tree that do not share a direct ascendant. A detailed description of the algorithm is provided, along with an extensive numerical analysis on both simulated and real datasets.

E0748: **K-bMOM: A robust K-means-type procedure with application to color quantization**
Presenter: **Edouard Genetay**, CREST-ENSAI, France
Co-authors: Adrien Saumard, Camille Saumard

Classical clustering methods, such as K-means, suffer from a lack of robustness with respect to outliers. We propose a robust version of K-means named K-bMOM, using bootstrap and median-of-means statistics, a strategy that has been recently put to emphasis for efficient, robust machine learning. The algorithm is iterative, in a Lloyd-type fashion. The performances of K-bMOM are theoretically and empirically shown. First, we give a theoretical majoration of the risk excess. Secondly, simulations show that K-bMOM converges rapidly along with the iteration steps, that it clearly outperforms K-means on corrupted or heavy-tailed data and that it is competitive with other robust approaches, such as K-median for instance. K-bMOM also provides interesting outcomes such as a robust and efficient initialisation procedure and outlier detection.

EO624 Room R14 RECENT DEVELOPMENTS IN QUANTILE REGRESSION Chair: Yudhie Andriyana

E0469: **Non-crossing spatial autoregressive quantile model applied to dengue fever incident in Bandung, Indonesia**
Presenter: **Yudhie Andriyana**, Universitas Padjadjaran, Indonesia

The existence of spatial dependence in a dataset needs to be accommodated by a proper model. One of the standard techniques often used is Spatial Autoregressive (SAR) model, which will be implemented to the incidence of dengue fever in Bandung, Indonesia. Dengue fever is an infectious disease that has impacts not only on a health aspect but also on social and economic aspects. Therefore, to prevent the disease, we need to control factors that influence the dengue's occurrence, considering the dependence between area. An interesting study is to know the model on the highest or the lowest risk of dengue fever, which cannot be solved by a classical regression technique. Therefore, we propose to use an existing technique called the quantile spatial autoregressive model. However, the technique is working with individual quantile objective function. In that case, it may lead to a crossings issue where the lower quantile levels may cross to the higher levels or vice versa. Hence, we propose a quantile technique to avoid such crossing problems.

E0679: **Improving linear quantile regression for replicated data**
Presenter: **Kausihik Jana**, Imperial College London, United Kingdom
Co-authors: Debasis Sengupta

An improvement of linear quantile regression is provided when there are a few distinct values of the covariates but many replicates. On can improve the asymptotic efficiency of the estimated regression coefficients by using suitable weights in quantile regression, or simply by using weighted least squares regression on the conditional sample quantiles. The asymptotic variances of the unweighted and weighted estimators coincide only in some restrictive special cases, e.g., when the density of the conditional response has identical values at the quantile of interest over the support of the covariate. The dominance of the weighted estimators is demonstrated in a simulation study and through the analysis of a data set on tropical cyclones.

E0754: **Semiparametric quantile regression**
Presenter: **Anneleen Verhasselt**, Hasselt University, Belgium

Quantile regression is an important tool in data analysis. Linear regression, or more generally, parametric quantile regression often imposes too restrictive assumptions. Nonparametric regression avoids making distributional assumptions but might have the disadvantage of not exploiting distributional modeling elements that might be brought in. A semiparametric approach towards estimating conditional quantile curves is proposed. It is based on a recently studied large family of asymmetric densities of which the location parameter is a quantile (and not a mean). Passing to conditional densities and exploiting local likelihood techniques in a multiparameter functional setting then leads to a semiparametric estimation procedure.

EO722 Room R15 ADVANCED TREE METHODS AND APPLICATIONS Chair: Rosaria Simone

E1114: **Trees and their building blocks**
Presenter: **Heidi Seibold**, Helmholtz AI, Germany

Tree-based methods continue to be popular in single trees and ensembles, and they come in various forms and implementations. We want to take

a bird's-eye view on tree algorithms and what they have in common: the general building blocks such as transformation, variable selection, and splitting. In particular, we will share what we have learned about trees in our work on the R package partykit.

E0383: **Ordinal forests: Prediction and covariate importance ranking with ordinal response variables**
Presenter: **Roman Hornung**, Institute for Medical Information Processing, Biometry and Epidemiology, University of Munich, Germany
The ordinal forest method is a random forest-type prediction method for ordinal response variables. The trees in ordinal forests are regression trees that use optimized score values in place of the ordered class values of the response variable. The optimization of the score values aims at maximizing the estimated prediction performance of the forest. Ordinal forests allow prediction using both low-dimensional and high-dimensional covariate data and can additionally be used to rank covariates with respect to their importance for prediction. An extensive comparison study using several real datasets and simulated data reveals that ordinal forests tend to outperform competitors in terms of prediction performance. Moreover, it is seen that the covariate importance measure currently used by ordinal forest discriminates influential covariates from noise covariates at least similarly well as the measures used by competitors. The rationale underlying ordinal forests of using optimized score values in place of the class values of the ordinal response variable is in principle applicable to any regression method beyond random forests for a continuous outcome that is considered in the ordinal forest method. While the original ordinal forest algorithm only also to perform class point predictions, a recent update allows predicting class probabilities.

E0901: **From unbiased MDI feature importance to explainable AI for trees**
Presenter: **Markus Loecher**, Berlin School of Economics and Law, Germany
Various recent attempts are unified to (i) improve the interpretability of tree-based models and (ii) debias the default variable-importance measure (MDI) in random forests. In particular, we demonstrate a common thread among the out-of-bag based bias correction methods and their connection to local explanation for trees. In addition, we point out a bias caused by the inclusion of inbag data in the newly developed SHAP values. Empirical and simulational studies indicate substantial improvements in the discriminative power of SHAP values when out-of-sample data are used instead.

EO550 Room R17 MODERN APPROACHES TO DIRECTIONAL DATA ANALYSIS Chair: Stefania Fensore

E0476: **Some advances on modal regression for circular data**
Presenter: **Rosa Crujeiras**, University of Santiago de Compostela, Spain
Co-authors: Maria Alonso-Pena
Modal regression is a quite effective alternative to classical regression methods when the conditional mean or median (or any other quantile) is not an adequate summary of the behavior of a response variable with respect a certain covariate. This usually happens when the conditional distribution shows asymmetry and/or when more than a single (conditional) mode is present in the data, leading in some scenarios to a more complex estimator (actually, a multifunction) than in mean or quantile regression. We will show how multimodal regression estimation can be accomplished for regression models involving circular variables (response and/or covariate), from a nonparametric perspective. The algorithms will be described in detailed and some examples on the cylinder and the torus will be shown to illustrate the methods.

E0633: **Circular order aggregation as a novel proposal to estimate sampling times in chronobiology**
Presenter: **Yolanda Larriba**, University of Valladolid, Spain
Co-authors: Cristina Rueda
Core clock circadian genes such as PERs, CRYs or ARNTL display clearly rhythmic expression patterns every 24 hours. The analysis of rhythmic genes, including pattern comparisons, amplitude or time point estimates, is crucial in chronobiology to understand biological functions correctly. However, daily expression data collection is expensive and may suppose a risk for health (e.g. human biopsies). Typically, in chronobiology, gene expression data are given from post-mortem specimens where expression data are measured at different times across several organs. This process involves a high uncertainty such as the exactly specimen time of death is usually unknown or not accurate; samples are taken from a large number of specimens; and because more than 30,000 expression data, including both rhythmic and non-rhythmic genes, are obtained from each organ simultaneously. Based on core clock genes, several directional approaches are evaluated, including circular principal component analysis, to solve temporal order estimation as a circular order aggregation problem. Specifically, statistical methods are applied to GTEx data collection, which contains more than 17,000 post-mortem RNA-Seq expression data across 54 human tissues. Preliminary results provide accurate sampling time estimates and enhance the knowledge of circadian biology.

E0930: **Depth-based classification of directional data**
Presenter: **Giuseppe Pandolfo**, University of Naples Federico II, Italy
Co-authors: Antonio D Ambrosio
A non-parametric procedure based on the concept angular depth function is developed for dealing with classification problems of objects in directional statistics. Several notions of depth for directional data are adopted: the angular simplicial, the angular Tukeys, the arc distance, the cosine distance and the chord distance depths. The proposed method is flexible and can be applied even in high-dimensional cases when a suitable notion of depth is adopted. Performances are investigated and compared by applying methods to different distributional settings through simulated and real datasets.

EO630 Room R21 CAUSAL SURVIVAL ANALYSIS Chair: Daniel Nevo

E0378: **Causal inference for semi-competing risks data**
Presenter: **Daniel Nevo**, Tel Aviv University, Israel
Co-authors: Malka Gorfine
An emerging challenge for time-to-event data is studying semi-competing risks, where two event times are of interest: the non-terminal event (e.g. disease diagnosis) time, and a terminal event (e.g. death) time. The non-terminal event is observed only if it precedes the terminal event, which may occur before or after the non-terminal event, leading to the latter being unobserved or even undefined. Studying treatment or intervention effects on the event times is complicated because, for some units, the non-terminal event time may occur only under one treatment value but not the other. We will present and discuss new estimands, based on time-fixed stratification of the population. These estimands correspond to the scientific questions of interest, as we will exemplify using a real-data example of the effect of the APOE gene on Alzheimer's disease and death. We will then present novel assumptions utilizing the time-to-event nature of the data. The new assumptions enable partial identifiability of causal effects of interest, namely bounds. We will also present and discuss a sensitivity analysis approach based on semi-parametric frailty models. Finally, we will present non-parametric and semi-parametric estimators for the causal estimands.

E0574: **New estimands for causal inference conditional on post-treatment variables**
Presenter: **Mats Stensrud**, Ecole polytechnique federale de Lausanne, Switzerland
Many studies aim to evaluate treatment effects on outcomes in individuals characterized by a particular post-treatment variable. For example, we may be interested in the effect of cancer therapies on quality of life, and quality of life is only well-defined in individuals who are alive. Similarly, we may be interested in the effect of vaccines on post-infections outcomes, which are only of interest in those individuals who become infected. In these settings, a naive contrast of outcomes conditional on the post-treatment variable does not have a causal interpretation, even in a randomized experiment. Therefore the effect in the principal stratum of those who would have the same value of the post-treatment variable regardless of

treatment, such as the survivor average causal effect, is often advocated for causal inference. Whereas this principal stratum effect is a well defined causal contrast, it cannot be identified without strong untestable assumptions, and its practical relevance is ambiguous because it is restricted to an unknown subpopulation of unknown size. We formulate alternative estimands, which allow us to define the conditional separable effects. We describe the causal interpretation of the conditional separable effects, e.g. in settings with truncation by death, and introduce three different estimators.

E0809: **Using negative controls to estimate the effect of treatment on survival when everyone is treated**
Presenter: **Ruth Keogh**, London School of Hygiene and Tropical Medicine, United Kingdom

Treatments are sometimes introduced for all patients in a particular cohort. When an entire cohort of patients receives a treatment, it is difficult to estimate its effect because there are no directly comparable untreated patients. The application that motivates this relates to a disease-modifying treatment in cystic fibrosis, ivacaftor, which has been available for everyone in the UK with a specific genotype since 2012. It is of interest to understand the causal effect of treatment on survival, which has not been assessed in randomized controlled trials, and also to project the potential long-term impact on life expectancy. To investigate this, we use observational longitudinal data on treatment use, genotype, survival and measures of patient health status from the UK Cystic Fibrosis Registry. Negative control outcomes observed in patients who do not receive the treatment are used to enable estimation of the causal effect of ivacaftor on survival, including patients eligible for the new treatment but before its availability (historical controls) and patients with an ineligible genotype (genotype controls). Causal diagrams and the potential outcomes framework are used to define the causal estimand of interest and to discuss the assumptions under which it can be estimated using the available control groups. We will discuss the use of different analysis models, including Cox regression and the additive hazards model.

EO157 Room R22 BAYESIAN MACHINE LEARNING Chair: Julyan Arbel

E0980: **Bayesian principles for learning machines**
Presenter: **Mohammad Emtiyaz Khan**, RIKEN Center for AI project, Japan

Humans and animals have a natural ability to learn and quickly adapt to their surroundings autonomously. How can we design machines that do the same? We will present Bayesian principles to bridge such gaps between humans and machines. We will show that a wide variety of machine-learning algorithms are instances of a single learning-rule derived from Bayesian principles. The rule unravels a dual-perspective yielding new mechanism for knowledge transfer in learning machines. It is claimed that Bayesian principles are indispensable for an AI that learns as efficiently as we do.

E0898: **Priors in Bayesian neural networks at the unit level**
Presenter: **Mariia Vladimirova**, Inria, France
Co-authors: Julyan Arbel

Neural networks (NNs), and their deep counterparts, have largely been used in many research areas such as image analysis, signal processing, or reinforcement learning, to name a few. The impressive performance provided by such machine learning approaches has greatly motivated research that aims at a better understanding of the driving mechanisms behind their effectiveness. In particular, the study of the NNs distributional properties through Bayesian analysis has recently gained much attention. We firstly describe the necessary notations and statistical background for Bayesian NNs. Then we consider its distributional properties and novel theoretical insight on distributions at the units level. Under the assumption of independent and normally distributed weights, we establish that the induced prior distribution on the units before and after activation becomes increasingly heavy-tailed with the depth of the layer.

E0905: **Interpreting a penalty as the influence of a Bayesian prior**
Presenter: **Pierre Wolinski**, Inria Grenoble, France
Co-authors: Guillaume Charpiat, Yann Ollivier

In machine learning, it is common to optimize the parameters of a probabilistic model, modulated by a somewhat ad hoc regularization term that penalizes some values of the parameters. Regularization terms appear naturally in Variational Inference (VI), a tractable way to approximate Bayesian posteriors: the loss to optimize contains a Kullback–Leibler divergence term between the approximate posterior and a Bayesian prior. We fully characterize which regularizers can arise this way, and provide a systematic way to compute the corresponding prior. This viewpoint also provides a prediction for useful values of the regularization factor in neural networks. We apply this framework to regularizers, such as L1 or group-Lasso.

EO614 Room R24 RECENT ADVANCES IN SCREENING DESIGNS Chair: Haralambos Evangelaras

E0816: **Fold-over weighing matrices for screening purposes and extensions with categorical data**
Presenter: **Stella Stylianou**, RMIT University, Australia
Co-authors: Stelios Georgiou

Screening is usually performed in the first stage of experimentation and aims in determining a small number of dominant factors among many potentially active factors which may affect a response. A class of three-level designs has been defined for screening in the presence of second-order effects using a variant of the coordinate exchange algorithm. Conference matrices have been used to construct definitive screening designs with good properties. Extensions with categorical factors were also introduced by constructing designs to accommodate some two-level qualitative factors using a DSD-augment and an ORTH-augment approach. We propose a method for the construction of efficient three-level screening designs based on weighing matrices and their complete fold-over. This method can be considered as a generalization previous methods. Orthogonal extensions of these designs, to include few categorical factors, are also considered. The approach is relatively straightforward, and no computer search is needed since our designs are constructed using known weighing matrices.

E0817: **Screening using new sensitivity metrics and design of experiments for combat simulations**
Presenter: **Stelios Georgiou**, RMIT University, Australia
Co-authors: Haydar Demirhan, Anil Dolgun, Andrew Gill, Stella Stylianou

The combined use of sensitivity metrics and design of experiments is introduced for the problem of identifying input factors which significantly influence the operational effectiveness ordering of a set of alternative systems that are modelled by a stochastic simulation. Two sensitivity metrics are compared and evaluated with the full factorial designs. The application to a combat simulation study provides encouraging results for the proposed methodology.

E0850: **Construction of optimal unbalanced two-level supersaturated designs**
Presenter: **Haralambos Evangelaras**, University of Piraeus, Greece

The $UE(s^2)$-optimal supersaturated designs with factors in two levels for factor screening have been previously introduced. These designs with k runs and n columns, $n > k$, do not possess factor balance like the traditional $E(s^2)$-optimal two-level supersaturated designs and is easier to be constructed for the most of the cases. We describe a method for constructing $k \times n$ $UE(s^2)$-optimal two-level supersaturated designs with additional desirable properties.

EG575 Room R20 CONTRIBUTIONS IN BAYESIAN MODELLING Chair: Daniel Henderson

E0307: **Subject-specific Bayesian hierarchical model for microbiome data analysis**
Presenter: **Matteo Pedone**, University of Florence, Italy
Co-authors: Francesco Stingo

Recent biomedical evidence suggests that the knowledge of microbiome composition and its function has a huge potential as a diagnostic tool. Motivated by the availability of microbiome abundance counts collected from different sources, clinical factors and diet-related covariates, the purpose is to explore associations between the microbial composition and the diet. Within the Dirichlet-multinomial regression framework, we propose a Bayesian hierarchical model that accounts for the complex structure of the interactions between diet and clinical factors. This leads to a high-dimensional framework, where sparsity is strongly induced via suitable priors and a thresholding mechanism. The model incorporates subject-specific regressions defined by coefficients that can vary flexibly with the covariates; the model effectively allows the effects of the covariates on the microbiome to be heterogeneous even when the sample size is small. An analysis of the microbiome abundance from patients affected by colorectal adenocarcinoma illustrates how the proposed approach can be used to determine the heterogeneous effects of diet and clinical factors on the microbiome.

E1109: **Modeling climatic and temporal influences on powerboat launches with relevance to recreational fisheries**
Presenter: **Ebenezer Afrifa-Yamoah**, Edith Cowan University, Australia

Digital camera monitoring data on recreational boating activity are often manually interpreted, and the reading cost can be expensive for multiple sites. Typically, this scheme is used along with other periodic boat-based surveys, and it is common practice that camera data between survey periods are not read, creating significant gaps in the time series. We predicted boating behaviour during these periods of non-observation using historical data and secondary variables to complete the time series data. Predictive models, built in a Bayesian regression modelling framework, were formulated to determine the temporal distribution of daily boating traffic at two ramps in Western Australia based on climatic variables (including temperature, humidity, wind speed and gust, sea level pressure and wind direction) and temporal classifications (including months, and day type). Constructed and reconstructed data generally aligned well with the observed data, with some temporal biases at the bulk and upper tail of the distributions. The 95% credible intervals of the reconstructed periods adequately captured the observed data at both locations. Data for the constructed periods depicted the general trends for the observed periods. Our results provide useful insights into using environmental factors to predict boating activity to fill in the gaps between survey years. This could assist in the ongoing monitoring and sustainable management of recreational fisheries.

E0794: **A Bayesian non-linear hierarchical framework for crop models based on big data outputs**
Presenter: **Muhammad Mahmudul Hasan**, Durham University, United Kingdom
Co-authors: Jonathan Cumming

Due to the increasing trend of world population, proper fertilization is very crucial for crop productivity to maintain the levels of food which will be required. We base our analysis on the big data output from the Environmental Policy Integrated Climate (EPIC) model, which provides time series output of the crop yield (among other outputs such as pollution indicators) in response to changes in inputs such as fertilizer levels, weather, and other environmental covariates. At the initial stage of our research, we investigate the results of the simulation of a full factorial design in nitrogen and phosphorus fertilizer levels for 3 different crops and crop rotations. We apply a non-linear Bayesian hierarchical model based on established yield models in order to make inferences about the response of crop yield with respect to fertilizers by using EPIC outputs. We use Markov Chain Monte Carlo to obtain samples from the posterior distributions, to validate and illustrate the results, and to carry out model selection. The results highlight a strong response of yield to nitrogen, but surprisingly a weak response to phosphorus for this particular simulator configuration and catchment.

CO710 Room R02 NONLINEAR, SEMI- AND NONPARAMETRIC PANEL DATA MODELING Chair: Markus Fritsch

C1152: **Flexible Bayesian modelling of treatment effects on panel outcomes**
Presenter: **Helga Wagner**, Johannes Kepler University, Austria

The estimation of the effects of a binary treatment on a continuous outcome observed over subsequent time periods is considered. We propose a new, flexible model that allows to separate longitudinal association of the outcomes from association due to endogeneity of treatment selection and employ this model to analyse the effects of a long maternity leave on earnings of Austrian mothers.

C0933: **R-Package pdynmc: GMM estimation of dynamic panel data models based on nonlinear moment conditions**
Presenter: **Joachim Schnurbus**, University of Passau, Germany
Co-authors: Markus Fritsch, Andrew Adrian Yu Pua

The R package pdynmc provides sufficient flexibility and precise control over the estimation and inference in linear dynamic panel data models. The package allows for the inclusion of nonlinear moment conditions and the use of iterated GMM; additionally, visualizations for data structure and estimation results are provided. The current implementation reflects recent developments in literature, uses sensible argument defaults, and aligns commercial and noncommercial estimation commands.

C0943: **Large sample properties of an IV estimator based on nonlinear moment conditions**
Presenter: **Markus Fritsch**, University of Passau, Germany
Co-authors: Andrew Adrian Yu Pua, Joachim Schnurbus

An instrumental variables (IV) estimator based is proposed on nonlinear (in parameters) moment conditions for estimating linear dynamic panel data models and derive the large sample properties of the estimator. We impose the following assumptions: (i) the only explanatory variable in the model is one lag of the dependent variable; (ii) the true lag parameter is smaller or equal to one in absolute value; (iii) the cross-section dimension is large; and (iv) the time series dimension is either fixed or large. Estimation of the lag parameter involves solving a quadratic equation, and we find that the lag parameter is point identified in the unit root case; otherwise, two distinct roots (solutions) result. We propose a selection rule that identifies the consistent root asymptotically in the latter case. We derive the asymptotic distribution of the estimator for the unit root case and the case when the absolute value of the lag parameter is smaller than one.

CO119 Room R03 TOPICS IN FINANCIAL ECONOMETRICS Chair: Leopold Soegner

C0912: **Bayesian reconciliation of the return predictability**
Presenter: **Borys Koval**, Vienna University of Economics and Business, Austria
Co-authors: Sylvia Fruehwirth-Schnatter, Leopold Soegner

A VAR for the returns, dividend growth, and dividend price ratio is estimated, where the Bayesian Control Function approach is applied to account for endogeneity. Motivated by financial literature we impose a stationarity condition on the auto-regressive dividend price ratio process by means of Bayesian priors. We develop two new priors, Jeffrey prior and prior based on frequentist bias-corrected 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 MSE, MAE, and credible interval coverage, the approach proposed in this article leads to superior performance relative to ordinary least squares estimation, a frequentist bias-corrected 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.

C1045: **G-identifiability for multivariate AR systems and mixed frequency data: REMIS for the unit root case**
Presenter: **Philipp Gersing**, Vienna University of Technology, Austria
Co-authors: Leopold Soegner, Manfred Deistler

The identification of the model parameters for data observed at mixed frequencies in a Johansen-type error correction model is investigated. Thus, the generic identifiability results for multivariate AR Systems and mixed frequency data are extended to the non-stationary case. We call this approach REMIS (REtrieval from MIxed Sampling frequency). For the blocked process of time-series observed, a canonical state-space representation is derived. By applying the projection approach recently developed in another paper of the authors, we obtain a system in prediction error form. For a regular high-frequency system, the blocked process is regular, and the system in prediction error form is minimal. Given the second moments of the variables observed under mixed frequency, the parameters of the high-frequency system are generically identified from these second moments. This result is established for slow stock variables as well as for variables obtained from a linear aggregation scheme.

C1129: **Triple the gamma: Achieving shrinkage and variable selection in TVP models**
Presenter: **Sylvia Fruehwirth-Schnatter**, WU Vienna University of Economics and Business, Austria
Co-authors: Peter Knaus, Annalisa Cadonna

Time-varying parameter (TVP) models are a popular tool for handling data with smoothly changing parameters. However, in situations with many parameters, the flexibility underlying these models may lead to overfitting models and, as a consequence, to a severe loss of statistical efficiency. This occurs, in particular, if only a few parameters are indeed time-varying, while the remaining ones are constant or even insignificant. As a remedy, hierarchical shrinkage priors have been introduced for TVP models to allow shrinkage both of the initial parameters as well as their variances toward zero. Various approaches to introducing shrinkage priors for TVP models are reviewed. Recently, the (hierarchical) triple Gamma prior has been introduced, which includes other popular shrinkage priors such as the double Gamma prior and the horseshoe prior as a special case. Efficient methods for MCMC inference are also discussed. The close resemblance of the triple Gamma prior with BM is investigated. For illustration, hierarchical shrinkage priors are applied to EU area inflation modelling based on the generalized Phillips curve, to a Cholesky stochastic volatility model, modelling multivariate financial time series of stock returns from the DAX, and to TVP-VAR-SV models, modelling multivariate macroeconomic time series. The results clearly indicate that shrinkage priors reduce the risk of overfitting and increase statistical efficiency in a TVP modelling framework.

CG094 Room R06 CONTRIBUTIONS IN FINANCIAL MARKETS Chair: Yingjie Dong

C0269: **Financial conditions, business cycle fluctuations and growth at risk**
Presenter: **Simone Manganelli**, European Central Bank, Germany
Co-authors: Andrea Falconio

The purpose is to study the macroeconomic consequences of financial shocks and increase in economic risk using a quantile vector autoregression. Financial shocks have a negative, but asymmetric impact on the real economy: they substantially increase growth at risk, but have limited impact on upside potential. The impact of financial shocks is explained away after controlling for economic risk (measured by the interquartile range). The effects are economically relevant. Bad economic environment, characterized by negative real and financial shocks, has a highly skewed impact on business cycle fluctuations, leading to a peak reduction of monthly industrial production by more than 2%. In comparison, positive real and financial shocks in a good economic environment have limited effect on upside potential of the economy.

C0652: **A COAALA copula for stock-bond return co-movement: Beware of the beast with four tails**
Presenter: **Anne-Florence Allard**, University of Bristol, United Kingdom
Co-authors: Hamza Hanbali, Kristien Smedts

The COAALA copula allows analysing financial market stability by studying comovement between stocks and government bonds using the information on both the global and the four local tail dependence measures (i.e. dependence between severe movements). Such an encompassing view on stock-bond co-movement has never been obtained convincingly before. Our contribution is twofold. First, we develop a novel copula function (COAALA) that is fully flexible. It accommodates the known features of stock-bond dependence. Also, it allows us to study the responses of each asset to a severe move of the other asset, taking into account potential asymmetries in the likelihood of such responses. This copula comes with closed-form expressions of dependence measures without the computational deficiencies of alternative models. Second, for a set of countries with very similar global dependence, the COAALA unravels major differences in local dependencies that hint at different stabilising abilities of bond markets across these countries.

C0717: **Comovement in cross-listed securities: Evidence from AH shares**
Presenter: **Yingjie Dong**, University of International Business and Economics, China
Co-authors: Wenxin Huang, Yiu-Kuen Tse

Dynamics of shares cross-listed in segmented markets are studied by investigating (1) the comovement between the level of prices of the same firm, and (2) the unobserved common factors in individual price differences (of the same firm) between markets. Using Chinese AH shares data, we employ the C-LASSO method to identify heterogeneous latent group patterns of H-share discounts and separate AH firms into two groups. These data-driven group patterns are associated with both firm and market characteristics. The results also identify both long-run and short-run common factors in AH price differences. The presence of the nonstationary (long-run) common factor suggests that, although AH shares trade on the same underlying stock, AH prices generally diverge away from the supposed comovement equilibrium. Thus, arbitrageurs cannot gain profit by exploiting deviations between AH prices. The detected short-run common factor suggests that AH price gap is also influenced by non-fundamental common shocks. The common divergence behavior of AH prices is related to several market economic variables.

CG020 Room R07 CONTRIBUTIONS IN FORECASTING II Chair: James Taylor

C0838: **Common dynamic factors for cryptocurrencies and multiple pair-trading statistical arbitrages**
Presenter: **Marco Patacca**, University of Verona, Italy
Co-authors: Gianna Figa Talamanca

Dynamic factor analysis is applied to model the joint behaviour of Bitcoin, Ethereum, Litecoin and Monero, as a representative basket of the cryptocurrencies asset class. The empirical results suggest that a model with two dynamic factors suitably describes the basket price. More precisely, we detect one integrated and one stationary factor until the end of August 2019 and two integrated factors afterwards. Based on this evidence, we define a multiple long-short trading strategy which proves profitable when the second factor is stationary.

C0663: **Forecast combination view of HAR model**
Presenter: **Andrey Vasnev**, University of Sydney, Australia
Co-authors: Adam Clements

The heterogeneous autoregressive (HAR) has become a popular model to predict realized volatility. It is simple and delivers good empirical results. Other modifications, e.g., HAR-Q model, were able to improve the results but only marginally. We take a step back and look at the HAR model

as a forecast combination model that combines three predictors: previous day realization (or random walk forecast), previous week average, and previous month average. When applying the OLS method to combine the predictors, the HAR model uses optimal weights that are known to be problematic in the forecast combination literature. In fact, the average forecast often outperforms the optimal combination in many empirical applications. We investigate the performance of the individual predictors. We then use equal weights instead of optimal weights and achieve up to 52% improvement as measured by the mean squared forecasting error. The combination framework opens a door of possibilities to construct the weights to achieve even more significant improvements. Smaller gains are observed in the context of multivariate HAR models when forecasting the covariance matrix of returns. However, these gains are meaningful, as simple combinations approaches avoid the estimation of large dimensional regression models.

C0190: **A simple model correction for modelling and forecasting (un)reliable realized volatility**

Presenter: **Rodrigo Hizmeri**, Lancaster University, United Kingdom

Co-authors: Marwan Izzeldin, Mike Tsionas

We propose a dilution bias correction approach to deal with the errors-in-variables problem observed in realized volatility (RV) measures. The absolute difference between daily and monthly RV is shown to be proportional to the relative magnitude of the measurement error. Therefore, in implementing the latter metric, and in allowing the daily autoregressive parameter to vary as a function of the error term, the result is more responsive forecasts with greater persistence (faster mean-reversion) when the measurement error is low (high). Empirical results indicate that our models outperform some of the most popular HAR and GARCH models across various forecasting horizons.

EO500 Room R12 CLUSTERING OF MULTIVARIATE DEPENDENT DATA Chair: Marta Nai Ruscone

E0546: **On correlation in a simple Gaussian sample that cannot be identified from data**
Presenter: **Christian Hennig**, University of Bologna, Italy

It is shown that $X_1,\ldots,X_n$ from an i.i.d. Gaussian sample cannot be distinguished based on observed data from Gaussian data for which the correlation $r(X_i,X_j)=\rho\neq 0\forall i,j$. This particularly implies that this violation of the i.i.d. assumption in Gaussian sampling cannot be checked, despite potentially having quite a strong impact on inference about the mean and variance parameters. A general definition of identifiability from data is introduced. How this is different from standard identifiability is discussed. Other parameters that are generally identifiable but not identifiable from data are cluster membership parameters in a spherical Gaussian fixed partition model as estimated by k-means clustering. There is, however, a different model for k-means where these parameters are identifiable from data.

E1072: **Clustering via copula-based dissimilarity measures**
Presenter: **Pier Giovanni Bissiri**, University of Bologna, Italy
Co-authors: Marta Nai Ruscone

A theoretical framework for clustering data is presented according to the dissimilarity behaviour as measured via a suitable copula-based coefficient and study its main properties. The coefficients are defined in terms of copulas, which may or may not be Gaussian. Applications to real data are used to illustrate the usefulness and importance of our proposal.

E0714: **Modeling dependency structures of UK SMEs: Combining clustering and spatial regression**
Presenter: **Antonia Gieschen**, The University of Edinburgh, United Kingdom
Co-authors: Raffaella Calabrese, Belen Martin-Barragan, Jonathan Ansell

Enabling small and medium-sized enterprises (SMEs) to prosper through access to finance is important for all economies. Whilst SMEs are diverse, they are linked by their location and other businesses they interact with. The aim is to investigate how these dependencies impact access to finance for SMEs. By combining spatial regression and clustering, we can explore the dependencies within spatially restricted networks. Using data on 3,227 UK SMEs spread across the UK, this method reveals evidence of a spillover effect that has a significant influence on their access to finance. Our method using clustering establishes networks of interconnected SMEs, which empirically highlight the need for a regional level policy which could aid SMEs in obtaining support for future funding.

E0189: **Semiparametric multinomial mixed-effects linear models: An expectation-maximization algorithm**
Presenter: **Chiara Masci**, Politecnico di Milano, Italy
Co-authors: Francesca Ieva, Anna Maria Paganoni

An expectation-maximization algorithm is proposed for semiparametric mixed-effects linear models dealing with a multinomial response. Multinomial Linear Mixed-effects Models (MLMMs) are often treated as multivariate models, where the integration issues typical of generalised linear mixed-effects models grow in complexity. In MLMMs, in order to obtain the marginal distribution of the response, random effects need to be integrated out. In a full parametric context, where random effects follow a multivariate normal distribution, this is often computationally infeasible. We propose a novel semiparametric approach in which random effects follow a multivariate discrete distribution with an a priori unknown number of support points, that is allowed to differ across categories. The advantage of this modelling is twofold: the discrete distribution on random effects allows, first, to express the marginal density as a weighted sum, avoiding numerical problems related to the integration and, second, to identify a latent structure at the highest level of the hierarchy, where groups are clustered into subpopulations. We show a simulation study and we apply the proposed algorithm to a real case study for predicting higher education student dropout, where students are nested within engineering degree programmes, comparing the results with the ones of a full parametric method.

E0297: **Vine copula mixture models and clustering for non-Gaussian data**
Presenter: **Ozge Sahin**, Technical University of Munich, Germany
Co-authors: Claudia Czado

The majority of finite mixture models suffer from not allowing asymmetric tail dependencies within components and not capturing non-elliptical clusters in clustering applications. Since vine copulas are very flexible in capturing these types of dependencies, we propose a novel vine copula mixture model for continuous data. We discuss the model selection and parameter estimation problems and further formulate a new model-based clustering algorithm. The use of vine copulas in clustering allows for a range of shapes and dependency structures for the clusters. The simulation experiments illustrate a significant gain in clustering accuracy when notably asymmetric tail dependencies or/and non-Gaussian margins within the components exist. The analysis of real data sets accompanies the proposed method. We show that the model-based clustering algorithm with vine copula mixture models outperforms the other model-based clustering techniques, especially for the non-Gaussian multivariate data.

EO179 Room R13 ML-ECO: MACHINE LEARNING AND STATISTICAL TECHNIQUES FOR ECONOMICS Chair: Vianney Perchet

E1013: **Economic mechanisms for crowdsourcing markets**
Presenter: **Alexey Drutsa**, Yandex, Russia

The current level of automation significantly affects the global economy, and the power of this influence continues to spread fast. On the one hand, this leads to a reduction in the need for labor and, thus, increases the unemployment rate. On the other hand, since machine learning technologies play a key role in automation, the need for collected and processed data is growing. In fact, data are now being transformed into a new kind of "oil" consumed by new "machines" (AI). Surprisingly, this need for data may be a clue to solve the problem of people unemployment and to balance the negative tendency in the global economy between humans and machines. In order to produce large amounts of data (for example, text entity recognition, object segmentation on a photo, etc.), machine learning-based products actively use crowdsourcing platforms (e.g., MTurk and Toloka) – platforms of a two-sided market between task requesters and their performers. The main feature of this market is that tasks are short, while performers can freely choose which tasks to execute. In particular, this is why this market is not similar to the classic labor market. We discuss how economic mechanisms can help crowdsourcing markets by overviewing main open research questions that arise while building a crowdsourcing platform. In particular, we consider the problem of "weak" matching: how to match performers to a ranking list of tasks taking into account the incentives of both sides of the market.

E1158: **Two-sided matching markets with correlated random preferences have few stable pairs**
Presenter: **Simon Mauras**, Universita de Paris, IRIF, France

Stable matching in a community consisting of N men and N women is a classical combinatorial problem that has been the subject of intense theoretical and empirical study since its introduction in 1962. We study the number of stable pairs, that is, the man/woman pairs that appear in some stable matching. We prove that if the preference lists on one side are generated at random using a given popularity model, the expected number of stable edges is bounded by $N\ln N+N$, matching the asymptotic value for uniform preference lists. If in addition that the popularity model is a geometric distribution, then the number of stable edges is $O(N)$ and the incentive to manipulate is limited. If in addition, the preference

lists on the other side are uniform, then the number of stable edges is asymptotically N up to lower-order terms: most participants have a unique stable partner, hence non-manipulability.

E1185: **Real-time optimisation for online learning in auctions**
Presenter: **Lorenzo Croissant**, Universite Paris Dauphine - PSL, France
Co-authors: Marc Abeille, Clement Calauzenes

In display advertising, a small group of sellers and bidders face each other in up to 10^{12} auctions a day. In this context, revenue maximisation via monopoly price learning is a high-value problem for sellers. By nature, these auctions are online and produce a very high-frequency stream of data. This results in a computational strain that requires algorithms to be real-time. Unfortunately, existing methods inherited from the batch setting suffer $O(\sqrt{t})$ time/memory complexity at each update, prohibiting their use. We provide the first algorithm for online learning of monopoly prices in online auctions whose update is constant in time and memory.

E1188: **An O(log log m) prophet inequality for subadditive combinatorial auctions**
Presenter: **Paul Duetting**, Google Research, Switzerland

Prophet inequalities compare the expected performance of an online algorithm for a stochastic optimization problem to the expected optimal solution in hindsight. They are a major alternative to classic worst-case competitive analysis, of particular importance in the design and analysis of simple (posted-price) incentive-compatible mechanisms with provable approximation guarantees. A central open problem in this area concerns subadditive combinatorial auctions. A number n of agents with subadditive valuation functions compete for the assignment of m items. The goal is to find an allocation of the items that maximize the total value of the assignment. The question is whether there exists a prophet inequality for this problem that significantly beats the best-known approximation factor of O(log m). We make major progress on this question by providing an $O(\log\log m)$ prophet inequality. The proof goes through a novel primal-dual approach. It is also constructive, resulting in an online policy that takes the form of static and anonymous item prices that can be computed in polynomial time given appropriate query access to the valuations. As an application of our approach, we construct a simple and incentive-compatible mechanism based on posted prices that achieves an $O(\log\log m)$ approximation to the optimal revenue for subadditive valuations under an item-independence assumption.

EO610 Room R14 MODEL SPECIFICATION TESTS Chair: Bojana Milosevic

E1001: **A new class of tests for the Pareto distribution based on the empirical characteristic function**
Presenter: **Jaco Visagie**, North-West University, South Africa
Co-authors: James Allison, Marius Smuts

The Pareto Type I distribution is a popular model in economics, finance and actuarial science, especially where phenomena characterised by heavy tails are studied. Due to the popularity of this distribution, goodness-of-fit tests have been developed to test the hypothesis that an observed dataset is compatible with the assumption of being realised from this distribution. Although tests exist for the Pareto distribution, they are few in number compared to those for other distributions such as, for example, the normal and exponential distributions. We propose a class of goodness-of-fit tests for the Pareto Type I distribution based on a characterisation involving the distribution of the sample minimum. The test is based on a weighted L2 norm between empirical characteristic function. A Monte Carlo study, involving the bootstrap, shows that the performance of the newly proposed tests compares favourably to existing tests for the Pareto distribution. A practical example relating to the earnings of professional golfers is included.

E0337: **New characterization based goodness-of-fit tests for randomly censored data**
Presenter: **Marija Cuparic**, University of Belgrade - Faculty of Mathematics, Serbia
Co-authors: Bojana Milosevic

Recently, the characterization based approach for the construction of goodness of fit tests has become popular. Most of the proposed tests have been designed for complete i.i.d. samples. We present the adaptation of the recently proposed exponentiality tests based on equidistribution-type characterizations for the case of randomly censored data. Their asymptotic properties are provided. We also present the results of a wide empirical power study, including the powers of several recent competitors. The results can be used as a benchmark for future tests proposed for this kind of data.

E0784: **Empirical process of concomitants for partly categorial data and applications in statistics**
Presenter: **Daniel Gaigall**, Leibniz University Hannover, Germany
Co-authors: Julian Gerstenberg

On the basis of independent and identically distributed bivariate random vectors, where the components are categorical and continuous variables, respectively, the related concomitants, also called induced order statistic, are considered. The main theoretical result is a functional central limit theorem for the empirical process of the concomitants in a triangular array setting. A natural application is hypothesis testing. An independence test and a two-sample test are investigated in detail. The fairly general setting enables limit results under local alternatives and bootstrap samples. Outputs of simulation studies confirm the theoretical findings.

E0952: **Change-point methods for estimating the proportion of alternative hypotheses**
Presenter: **Anica Kostic**, London School of Economics and Political Science, United Kingdom
Co-authors: Piotr Fryzlewicz

The problem of estimating the number of alternative hypotheses among a large number of independently tested hypotheses is considered. The method utilizes the idea of separating small, possibly alternative p-values, from the null p-values, by estimating the change-point location in the sequence of sorted p-values. This can be applied to the problem of fitting a high-dimensional mean-shift model with aligned change-points, specifically for estimating the proportion of components that have changed at a given time. We illustrate the use of this method on DNA copy number variants data to describe the extent of abnormal numbers of copies across the population.

E0594: **Tests for heteroskedasticity in transformation models**
Presenter: **Charl Pretorius**, North-West University, South Africa
Co-authors: Simos Meintanis, Marie Huskova

A model is considered whereby a given response variable Y following a transformation $\mathcal{T}(Y)$ satisfies some classical regression equation. In this transformation model, the form of the transformation is specified analytically but incorporates an unknown transformation parameter that needs to be estimated. We develop testing procedures for the null hypothesis of homoskedasticity for versions of this model where the regression function is considered to be either known or unknown. The test statistics are formulated on the basis of Fourier-type conditional contrasts of a variance computed under the null hypothesis against the same quantity computed under alternatives. The limit null distribution of the test statistic is studied, as well as the behaviour of the test criterion under alternatives. Since the limit null distribution is complicated and involves many unknown quantities, a bootstrap scheme is suggested in order to carry out the test procedures. Monte Carlo results, which illustrate the finite-sample properties of the new procedures, will also be presented.

EO225 Room R15 TOPICS ON HIGH-DIMENSIONAL METHODOLOGY Chair: Eugen Pircalabelu

E0410: **Lasso principal support vector machines for sufficient dimension reduction**
Presenter: **Andreas Artemiou**, Cardiff University, United Kingdom
Co-authors: Eugen Pircalabelu

A new method is developed for performing dimension reduction for high-dimensional settings. The proposed procedure is based on a principal support vector machine framework where principal projections are used to overcome the non-invertibility of the covariance matrix. Using a series of equivalences, we show that one can accurately recover the central subspace using a projection on a lower-dimensional subspace and then apply an l1 penalization strategy to obtain sparse estimators of the sufficient directions. Based next on a desparsified estimator, we provide an inferential procedure for high-dimensional models that allows testing for the importance of variables in determining the sufficient direction. Theoretical properties of the methodology are illustrated, and simulated and real data experiments demonstrate computational advantages.

E0534: **Asymptotic mean squared error of model-averaged M-estimators in high dimensions**
Presenter: **Jing Zhou**, KU Leuven, Belgium
Co-authors: Gerda Claeskens, Jelena Bradic

The focus is on the asymptotic mean square error (AMSE) of the estimators of the coefficient vector in a sparse high dimensional linear model where sample size n and the number of parameters p increase such that $n/p \to \delta \in (0,1)$. The approximate message passing (AMP) algorithm is considered for obtaining the AMSE expressions of the l_1-regularized M-estimators when $n, p \to \infty$. Instead of using a single M-estimator, we consider weighting multiple estimators by model averaging. We investigate the convergence of multiple correlated M-estimators. We further obtain the limit AMSE expression of the l_1-regularized M-estimators incorporating the selection uncertainty of the nonzero components of the coefficient vector due to regularization. An analytical expression of an AMSE-type weight choice for the model-averaged estimators is derived by minimizing the AMSE expression. For practical use, we construct a Stein-type estimator of the AMSE expression. We further use the results to construct componentwise confidence intervals and perform hypothesis testing on the model parameter.

E0619: **Corrected information criterion for coefficient selection and estimation in structured sparse linear regression**
Presenter: **Bastien Marquis**, Universita libre de Bruxelles, Belgium
Co-authors: Maarten Jansen

In high-dimensional linear regression, when the vector of coefficients is sparse, regularisation is widely used to obtain an estimate. In particular, ℓ_1-regularisation has many attractive qualities as it allows a selection of nonzero coefficients, in addition, to be computationally efficient; however, its solutions are shrunk versions of the ordinary least squares estimator. This can lead to a bias amongst the large coefficients but also results in an overestimation of the model size from the optimisation of an information criterion. To prevent these effects, we propose to use ℓ_1-regularisation as a method to select nonzero coefficients while using a least-squares projection for the estimation of the selection, avoiding the shrinkage this way. Then the optimal balance between the sum of residual squares and the regularisation should shift towards smaller models. This requires a correction of the expression of the information criterion. Looking into the difference between the Prediction Error and the expected Mallows's Cp, a corrected Mallows's Cp is developed for multiple linear regression. The correction is further analysed in structured models; in particular, group selection is considered.

E1149: **Distributed high-dimensional estimation and inference for precision matrices**
Presenter: **Ensiyeh Nezakati Rezazadeh**, Universita catholique de Louvain (Belgium), Belgium
Co-authors: Eugen Pircalabelu

Precision matrix estimation plays an important role in statistical and machine learning framework. When the sample size n or the dimension of the dataset is large, estimation of the precision matrix using one single computer is computationally challenging. An attractive approach for down-scaling the problem size is splitting dataset to subsets and fit models using distributed algorithms. The dataset can be split either based on the observations or based on the variables. Much of the attention has been focused on splitting the observations into K independent sub-samples that are analyzed in parallel. While splitting on the p variables is more effective when $p >> n$. We present a lasso-type distributed estimator of the precision matrix for high-dimensional Gaussian graphical models by implementing a partitioning procedure on both the observations and the features at the same time. A new combined estimator in each sub-sample is introduced by 'glueing' together the local estimators of every subset of variables and the cross estimators of every two subsets of variables. We use the asymptotic distribution of these combined estimators and introduce a confidence distribution-based method to aggregate the estimators and build a new final estimator. Theoretical properties are investigated, and a simulation study and a real data example are used to assess the performance of this estimator.

EO073 Room R17 TOPICS IN TIME SERIES Chair: Konstantinos Fokianos

E0467: **R-estimators in GARCH models: Asymptotics, applications and bootstrapping**
Presenter: **Kanchan Mukherjee**, Lancaster University, United Kingdom

The quasi-maximum likelihood estimation is a commonly-used method for estimating the GARCH parameters. However, such estimators are sensitive to outliers, and their asymptotic normality is proved under the finite fourth-moment assumption on the underlying error distribution. We propose a novel class of estimators of the GARCH parameters based on ranks of the residuals, called R-estimators, with the property that they are asymptotically normal under the existence of a finite $2+\delta$ moment of the errors and are highly efficient. We also consider the weighted bootstrap approximation of the finite sample distributions of the R-estimators. We propose fast algorithms for computing the R-estimators, and their bootstrap replicates. Both real data analysis and simulations show the superior performance of the proposed estimators under the heavy-tailed distributions and the excellent coverage rates of the weighted bootstrap approximations.

E0801: **Fast parameter and confidence interval estimation for Hidden Markov Models using template model builder**
Presenter: **Timothee Bacri**, University of Bergen, Norway
Co-authors: Jan Bulla, Geir Berentsen

Hidden Markov Models (HMMs) are a class of models widely used in speech recognition and can help other fields to model data such as phylogenetic trees or rainfall occurrence. There are straightforward ways to compute maximum likelihood estimates (MLEs) of their parameters. However, obtaining confidence intervals usually is more difficult. In addition, computing MLEs can be time-consuming for large datasets and complex models. We present a way to speed up core computational procedures for maximum likelihood estimation by up to 50 times compared to common optimization approaches. At the same time, we retrieve reliable estimates of standard errors within our framework. Firstly, we investigate how to optimize a Poisson HMM with the `TMB` package in R and how to retrieve confidence intervals. Secondly, we compare different optimizers (such as, e.g. `nlminb`) and minimize the negative log-likelihood directly on different datasets.

E0414: **Testing independence under local stationarity via the local empirical characteristic function**
Presenter: **Guy-Niklas Brunotte**, Otto-Friedrich-Universitat Bamberg, Germany

Locally stationary processes are non-stationary processes which can be locally approximated by stationary processes (so-called companion processes) allowing us to handle them appropriately by using tools for stationary processes. Kac's theorem shows that independence between random variables can be expressed essentially by their characteristic functions. This is the basic idea behind the construction of a consistent level-alpha test for independence via a local version of the empirical characteristic functions. A global test for pairwise independence of companion processes

at the same points in time is introduced which evaluate the L2-distance between the joint local empirical characteristic function of two locally stationary processes and the product of the local empirical characteristic functions of these processes.

E0549: **Time series parameter estimation for ocean wave models**
Presenter: **Jake Grainger**, Lancaster University, United Kingdom
Co-authors: Adam Sykulski, Philip Jonathan, Kevin Ewans

Understanding the behaviour of wind-generated ocean waves is important to many offshore and coastal engineering activities. Estimating models for the frequency domain behaviour of wind-generated wave time series has received considerable attention in the oceanographic literature. Typically, parametric spectral forms (such as JONSWAP) are fitted to periodograms calculated from observed time series, using least-squares techniques. We demonstrate, both by simulation and by theoretical reasoning, that some spectral model parameters are difficult to estimate using this approach. In contrast, using de-biased Whittle likelihood-based inference, we obtain more accurate and precise parameter estimates. To improve the computational efficiency of de-biased Whittle inference, we introduce a new technique to calculate likelihood derivatives and to approximate the variance of the resulting spectral estimator.

E0722: **Mixtures of nonlinear Poisson autoregressions**
Presenter: **Konstantinos Fokianos**, University of Cyprus, Cyprus

Non-linear infinite order Markov switching integer-valued ARCH models for count timeseries data are studied. Markov switching models take into account complex dynamics and can deal with several stylistic facts of count data including proper modelingof non-linearities, overdispersion and outliers. We study structural properties of those models. Under mild conditions, we prove consistency and asymptotic normality of the maximum likelihood estimator for the case of finite order autoregression. In addition, we give conditions which imply that the marginal likelihood ratio test, for testing the number of regimes,converges to a Gaussian process. This result enable us to prove that the BICprovides a consistent estimator for selecting the true number of regimes. A real data example illustrates the methodology and compares this approach with alternative methods.

EO259 Room R20 RECENT DEVELOPMENTS IN MULTIVARIATE DATA ANALYSIS Chair: Klaus Nordhausen

E0365: **Notion of information and independent component analysis**
Presenter: **Una Radojicic**, Technical University of Vienna, Austria
Co-authors: Klaus Nordhausen, Hannu Oja

In the engineering literature, independent component analysis is often described as a search for the uncorrelated linear combinations of the original variables that maximize non-Gaussianity. The estimation procedure usually has two steps. First, the vector of principal components is found, and the components are standardized to have zero means and unit variances. Second, the vector is further rotated so that the new components maximize a selected measure of non-Gaussianity. It is then argued that the components obtained in this way are made as independent as possible or that they display the components with maximal information. The information measures and measures of non-Gaussianity, including third and fourth cumulants are generally used as projection indices in the projection pursuit approach for the independent component analysis. We discuss and clarify the vague connections between non-Gaussianity, independence and notions of information in the context of the independent component analysis by discussing partial orderings and various measures of information for continuous univariate random variables with special roles of Gaussian and uniform distributions.

E0505: **Lassoing eigenvalues**
Presenter: **Mengxi Yi**, University of International Business and Economics, China

The properties of penalized sample covariance matrices depend on the choice of the penalty function. We will introduce a class of non-smooth penalty functions for the sample covariance matrix, and demonstrate how this method results in a grouping of the estimated eigenvalues. We refer to this method as lassoing eigenvalues or as the elasso.

E0586: **Radial growth models for geometric objects**
Presenter: **John Kent**, University of Leeds, United Kingdom

Consider an object in two or three dimensions evolving through time. For example, the head of a child changes in shape and size as it grows into an adult. Various radial growth models have been proposed in the literature to give a simplified description of growth, notably the cardioid strain models of Todd and Mark. These models involve a "seed", typically interior to the object, such that (a) growth occurs along rays emanating from the seed, and (b) the magnitude of growth depends on the distance from the seed. Modified versions of these models, based on linear regression and directional statistics, are developed here to facilitate the estimation of the growth parameters. To fit such models in practice, suppose that a set of landmarks has been identified on the object and that the object has been observed at two distinct ages. The hardest part of fitting these growth models is the estimation of the seeds. For planar objects it is possible, given the seeds, to write the maximized log-likelihood over the remaining parameters in closed form. Hence it is possible to examine the goodness-of-fit of the model, especially how well-determined the seeds are, through a numerical grid search in four dimensions. Some examples will be given to illustrate the strengths and limitations of these radial growth models.

E0773: **Revisiting estimation methods for spatial econometric interaction models**
Presenter: **Lukas Dargel**, Universite Toulouse 1 Capitole, France
Co-authors: Thibault Laurent

Spatial interactions describe phenomena as diverse as international trade flows, migration flows, or passenger flows in public transport. Explaining the main causes of such origin-destination flows is, therefore, a goal that unites empirical researchers and practitioners from various industries and scientific disciplines. The best-known interaction model is the gravity model, but it leads to inconsistent parameter estimates if there is some dependence between the flows. Techniques from the field of spatial econometrics solve this problem by including explicit measures of network dependence in the model. However, the parameters of this extended model cannot be estimated by ordinary least squares, and we have to adopt more sophisticated estimation methods, such as Maximum Likelihood (MLE), spatial two-stage least squares (S2SLS) or Bayesian Markov chain Monte Carlo (MCMC). We develop improved calculations for these three estimators, and our R package makes them operational for a broader public. Our simulations show that we can estimate the model parameters consistently and that the MLE is superior to S2SLS and MCMC estimation in terms of computational performance.

E1012: **Multivariate functional outlier detection using invariant coordinate selection**
Presenter: **Anne Ruiz-Gazen**, Toulouse School of Economics, France
Co-authors: Aurore Archimbaud, Feriel Boulfani, Xavier Gendre, Klaus Nordhausen, Joni Virta

Invariant Coordinate Selection (ICS) is an unsupervised multivariate method based on the joint diagonalization of two scatter matrices. It is a dimension reduction method which leads to outlyingness scores helpful for outlier detection in a multivariate context. Nowadays, more and more data sets are of multivariate functional nature, and various possibilities can be considered to extend the ICS outlier detection method to the multivariate functional framework. As usual in functional data analysis, we consider that the multivariate measurements correspond to functions observed on a discrete set of points in their domain. One possible extension of ICS consists in calculating for each component of the vector of curves, a functional approximation of the observed curves using some suitable basis and a finite number of basis vectors. ICS is then implemented on the stacked vector of the coordinates of the component functions in the basis of interest. Another possible extension is to calculate ICS scores at each domain point and derive some global outlyingness measurements over the domain. The two approaches are compared on several real data

examples, including some flight monitoring data from the aeronautics industry.

EO590 Room R21 INFERENCE IN SURVIVAL MODELS Chair: Giuliana Cortese

E0806: **Empirical likelihood inference to compare t-year absolute risks with right censored competing risks data**
Presenter: **Paul Blanche**, University of Copenhagen, Denmark

The t-year absolute risk, also called the cumulative incidence function at time t, is an interesting quantity routinely estimated in the competing risks setting. It is often estimated with the non-parametric Aalen-Johansen estimator. This estimator handles right-censored data and has desirable large sample properties, as it is the non-Parametric maximum likelihood estimator (NPMLE). Inference for comparing of absolute risks, via risk difference or risk ratios, can therefore be done via usual asymptotic normal approximations and the use of the delta-method. However, the small sample performance of this approach can be modest. Especially (i) coverage of confidence intervals can be poor, and (ii) inference using risk ratios and risk difference can lead to inconsistent conclusions, in terms of significant differences. We, therefore, introduce an empirical likelihood ratio based inference as an alternative. One advantage is that it always leads to consistent conclusions when comparing absolute risks via either risk ratios or risk differences, in terms of significance. Simulation results also suggest that small sample inference using this approach can be more accurate. We present how to compute the new confidence intervals and p-values. Novel technical results include formulas and algorithms to compute constrained NPMLE, from which likelihood ratios and inference procedures are derived. Examples using medical data are provided.

E0366: **Adjusted score functions to solve monotone likelihood problems in the Cox regression model**
Presenter: **Euloge Clovis Kenne Pagui**, University of Padova, Italy

Standard inference procedures for the Cox model involve maximizing the partial likelihood function. A phenomenon well known as monotone likelihood might be observed when fitting the partial hazard model to particular data sets. Monotone likelihood mainly occurs in samples with substantial censoring of survival times and is associated with categorical covariates. In particular, and more frequently, it usually happens when one level of a categorical covariate has just experienced censoring times. One way to overcome this problem is to use of adjusted partial likelihood score aiming at mean bias reduction. The procedure is effective in preventing infinite estimates. As an alternative solution, we propose an approach based on the adjusted score function for median bias reduction. This procedure also solves the infinite estimate problem and has an additional advantage of being invariant under component-wise reparameterizations. This latter aspect is fundamental under the Cox model since hazards ratio interpretation is obtained by exponentiating parameter estimates. Extensive numerical studies of the proposed method suggest better inference properties than those of the mean bias reduction. A real data application related to a melanoma skin data set is used as an illustration for a comparison basis of the methods.

E0992: **Personalised screening schedules for optimal prevention of cardiovascular disease**
Presenter: **Francesca Gasperoni**, MRC Biostatistics Unit, University of Cambridge, United Kingdom
Co-authors: Paul Newcombe, Chris Jackson, Angela Wood, Jessica Barrett

Cardiovascular disease (CVD) population screening strategies aim to identify and treat people at high risk of CVD. Current UK guidelines recommend screening adults over 40 years old every 5 years and prescribing statins for those with a predicted 10-year CVD risk greater than 10%. We propose an incremental net benefit function to investigate a personalised screening schedule, considering personal CVD risk profile. This function is composed of benefit (event-free life years) and costs (of statins and visits provided by the health services). The prescription of statins is assumed to start at the first visit after the 5-year CVD risk exceeds the 5% threshold. To assess this risk by adjusting for time-varying endogenous covariates, we use a two-stage dynamic landmark model. The first stage consists in fitting at each landmark age (i.e., 40,45,..,80 years) a multivariate linear mixed effect model with random intercepts and slopes. The second stage consists in predicting the CVD risk through a Cox model, adjusted for the risk factor values estimated at stage one. We apply the proposed model to data from the Clinical Practice Research Datalink (CPRD), comprising primary care Electronic Health Records from the UK. From preliminary analyses, baseline characteristics play a significant role in the optimal schedule. In particular, people labelled as high-risk seem to require more frequent visits, while low-risk people seem to require visits less frequently than every 5 years.

E0848: **A screening selection method for ultrahigh-dimensional survival data**
Presenter: **Sara Milito**, University of Salerno, Italy
Co-authors: Marialuisa Restaino, Francesco Giordano

With the recent explosion of computing power, modern studies in many areas, such as medicine, biosciences, demography, economy, generate a large amount of survival data. Selecting significant variables plays a crucial role in model building, and it becomes particularly challenging in an ultra-high dimensional setting where the dimension of covariates can be much larger than the sample size. In this context, since it is crucial to identify the variables that influence the survival time, the main aim is to reduce the dimensionality of the problem. Extensive work has been carried out for variable screening, that is the process of filtering out most of the irrelevant variables. In contrast, all relevant variables survive with probability tending to 1. Most of the methods available in the literature for survival models consider a conditional estimate of the survival function, using the Kaplan and Meier estimator (KM), in order to capture the impact of the variables one by one. This estimator has some disadvantages, especially with continuous covariates, since its formulation does not involve covariates. It is possible to estimate the covariate's effect directly on survival function using a different approach, not involving the KM estimator. Therefore, we aim to suggest a new procedure that overcomes the disadvantages of KM estimator. Some simulation results show that our proposed method performs satisfactorily.

E0379: **Survival analysis in two-stage randomized clinical designs using mixture distributions**
Presenter: **Giovanna Ranzato**, University of Padova, Italy
Co-authors: Giuliana Cortese

In many clinical designs, patients are treated by different combinations of therapies. In a two-stage design specifically, they are randomized to a primary therapy and eventually to a secondary therapy, depending on disease remission and patients' consent. Since the aim is to achieve the largest overall clinical benefit, the total effect of different combinations of first-stage and second-stage treatments on a survival outcome is of great interest. We propose a parametric estimator of the combined survival function for two-stage treatment strategies, using mixture distributions to model the possibly right-censored survival time. Observations are allowed to be censored in both stages; the first-stage duration is also modeled. The proposed parametric approach is particularly useful to investigate possible dissimilarities across strategies since parameters related to the first stage or second stage outcomes can be easily tested under a well-known likelihood framework. Simulation studies show a good performance of our estimator and an application to a two-stage randomized study on leukemia patients reveals that the procedure is easy to implement in practice.

EO053 Room R22 TRACK: BAYESIAN SEMI- AND NON-PARAMETRIC METHODS I Chair: Bernardo Nipoti

E0832: **Model misspecification and familial null hypotheses**
Presenter: **Catherine Forbes**, Monash University, Australia

One of the most basic questions in Statistics is "Are the centers of two distributions the same?". The question is answered formally through a hypothesis test with a null of "no difference", whether the analysis be classical or Bayesian. Traditional formulations of the problem rely on the belief that the models are perfectly specified. Robustness to violations of assumptions is typically studied under conditions that do not change the validity of the null hypothesis (e.g., symmetric contaminations). The actual deficiencies in model and data are likely to change a true null into a false statement. These imperfections suggest the use of flexible nonparametric models for the two (possibly paired) distributions and also suggest

consideration of a family of measures of center (e.g., for real-valued data, those based upon Huber's loss function). Each measure of the center generates a testing problem. The resulting family of null hypotheses constitutes a familial null hypothesis. Profile methods replace the original question with "Is there a measure of center in the family for which the centers of the two distributions are the same?"

E0239: **Bayesian joint modelling of recurrence and survival: A conditional approach**
Presenter: **Willem van den Boom**, National University of Singapore, Singapore
Co-authors: Maria De Iorio, Marta Tallarita

Recurrent event processes describe the repetition of an event over time. A recurrent event process is often terminated or censored by another event with dependence between the termination time and recurrence process. For instance, recurrent disease events might be terminated by death, while frailty might affect both disease recurrence and survival. As such, it is important to model the recurrent event process and the termination time process jointly to better capture the dependency between them. We propose a model in which the number of gap times, i.e. the time between two consecutive recurrent events, before the terminal event occurs is a random variable of interest. Then, conditionally on the number of recurrent events before the termination event, we specify a joint distribution for the gap times and the survival time. This novel conditional approach induces dependence between the recurrence and survival process. Additional dependence between recurrence and survival is introduced by a joint distribution on their respective frailty terms. A non-parametric random effects distribution for the frailty terms accommodates population heterogeneity and allows for data-driven clustering of the subjects. Posterior inference is performed through a tailor-made Gibbs sampler strategy involving a reversible jump step and slice sampling.

E0424: **Transport-based measure of dependence for Bayesian nonparametric models**
Presenter: **Marta Catalano**, Bocconi University, Italy
Co-authors: Antonio Lijoi, Igor Pruenster

Dependent random measures are a prominent tool for performing Bayesian nonparametric inference across multiple populations. The borrowing of strength across different samples is regulated by the dependence structure of the random measures, with complete dependence corresponding to the maximal share of information and fully exchangeable observations. For a substantial prior elicitation, it is crucial to quantify the dependence in terms of the hyperparameters of the models. State-of-the-art methods partially achieve this through the expression of the pairwise linear correlation. We propose the first non-linear measure of dependence for random measures. Dependence is characterized in terms of distance from exchangeability through a suitable transport metric on vectors of random measures. This intuitive definition extends naturally to an arbitrary number of samples, and it is analytically tractable on noteworthy models in the literature.

E0451: **Dependent prior processes for panel count data**
Presenter: **Beatrice Franzolini**, Bocconi University, Italy
Co-authors: Antonio Lijoi, Igor Pruenster

Panel count data occur in observational studies and clinical trials that concern recurrent events, where for each subject cumulative counts are recorded at discrete time points. Both the times points and the cumulative counts are realizations of point processes, namely the observation process and the event process. Even though assuming independence between the two simplifies the inference procedure, the assumption is not realistic in many applications. Prior information on the relation between counts and observation times is often available. We propose a class of Bayesian nonparametric priors over the observation and the event processes that allows for dependence between them, incorporating prior information on the positive association between frequency of observation and counts. The priors are defined modeling the intensities of the two processes through mixtures with respect to GM-dependent completely random measures. We investigate prior and posterior distributional properties of the model and develop a Markov Chain Monte Carlo algorithm to perform posterior inference. The merits of the proposal are further discussed through illustrative examples.

E1187: **Variational Bayes for model averaging for multivariate models using compositional predictors**
Presenter: **Alex Lewin**, London School of Hygiene and Tropical Medicine, United Kingdom

High-throughput technology for molecular biomarkers produces multivariate data exhibiting strong correlation structures and thus should be analysed in an integrated manner. Bayesian models are strongly suited to this aim. A particular case of interest is microbiome data, which is inherently compositional, and thus imposes a constraint on model space. A Bayesian model is presented for multivariate analysis of high-dimensional outcomes and high-dimensional predictors, including compositional predictors. The model includes sparsity in feature selection for predictors and covariance selection. A model averaging approach is taken to ensure a robust selection of predictors. A hybrid Variational Bayes - Monte Carlo computational approach is used for the compositional data updates.

EO634 Room R24 RECENT ADVANCES ON DESIGN OF EXPERIMENTS Chair: Victor Casero-Alonso

E0466: **Evolution study of related characteristics of a population from the point of view of design**
Presenter: **Juan M Rodriguez-Diaz**, University of Salamanca, Spain
Co-authors: Rosa Eva Pruneda, Maria Mercedes Rodriguez-Hernandez

The evolution of characteristics from a population that may be correlated will be studied from the point of view of design. Depending on the type of the observations, the corresponding tests needed for the follow-up of the variables could be expensive or time-demanding. Thus, the aim is usually to decide the most convenient moments where a limited number of tests should be done to obtain the maximum information. Using an optimal design of experiments approach, a general method for selection of the best temporal points will be developed. An application to the study of two variables related to the capacity of resolution of mathematical problems for primary-school students will be shown, employing different evolution models.

E0490: **Optimal designs for Antoine's equation**
Presenter: **Carlos de la Calle-Arroyo**, Universidad de Castilla-La Mancha, Spain
Co-authors: Jesus Lopez-Fidalgo, Licesio Rodriguez-Aragon

The influence of temperature in the vapour pressure of liquid or gas is derived from a class of semi-empiric equations known as Antoine's Equation. It represents the relationship between vapour pressure and temperature for distillation or pharmacological processes, in which a precise estimation of the parameters is required. The model has three unknown parameters for which, as a non-linear model, previous best guesses are required. These best guesses vary for different pure substances, and between the state of the substance, which establishes the space of the design for practical matters. To improve the accuracy of the experimental experiences with this model, optimal designs have been proposed. The analytical expression for the D-optimal design is included. Optimal designs have been calculated numerically for Ds-optimality, which has focused on estimating with least variance a subset of the parameters, A-optimality, that minimises the average of the variances, and I-optimality, to minimize the variance of the prediction over an interest region. Comparisons between some common designs and optimal designs are shown, to benchmark usual choices in this kind of experiments. Also, a class of designs with a compromise between both has been computed. An online tool to calculate Antoine's optimal designs for the considered criteria has been developed.

E0651: **Development of robust designs for accelerated failure time models with Type I censoring**
Presenter: **Irene Garcia-Camacha Gutierrez**, University of Castilla-La Mancha, Spain
Co-authors: M Jesus Rivas-Lopez, Raul Martin-Martin

Accelerated Failure Time (AFT) models are commonly used in the field of manufacturing, but they are more and more frequent for modeling clinical trial data. These models are defined through the survival function of the time-to-event variable, T. The construction of robust designs for AFT models is considered, with the possibility that the Acceleration Factor (AF) is misspecified when the variance of T is known. In particular, the "true" AF is allowed to vary over a neighbourhood of possible functions, $AF(\mathbf{x},\theta) = \exp(\theta^T\mathbf{x} + g_n(\mathbf{x}))$, for some unknown perturbation function g_n. Thus the efficiency of a design cannot be assessed through the covariance matrix of the Maximum Likelihood Estimator (MLE). Still, the estimate is subject to both "bias error" due to the inadequacy of the model as well as "variance error" due to sampling. The asymptotic mean squared error matrix (MSE) of the parameter estimates is obtained for right-censored observations. D- and I-optimal robust designs are derived from the above result considering the log-logistic distribution for illustration.

E0740: **Subsampling from big datasets through optimal design**
Presenter: **Chiara Tommasi**, University of Milan, Italy
Co-authors: Laura Deldossi

Big Data are huge amounts of digital information that rarely result from properly planned surveys; as a consequence, they often contain redundant data. A Big Dataset is herein conceived as a finite population generated by a super-population model. When the aim is to answer a particular question of interest, we suggest selecting a subsample of observations that contains the majority of the information to achieve this inferential goal. The selection methods driven by the theory of optimal design incorporate inferential purposes and thus perform better than standard sampling schemes.

E0700: **Optimal compound designs for the problem of uncertainty in probability distributions**
Presenter: **Sergio Pozuelo Campos**, University of Castilla-La Mancha, Spain
Co-authors: Mariano Amo-Salas, Victor Casero-Alonso

The uncertainty in the probability distribution of the response variable may have an important influence on the calculation of the optimal design of an experiment. The main goal is to study the effect of this uncertainty on the optimal design and to propose robust alternatives through compound optimal designs for generalized regression models considering usual probability distributions of the exponential family.

EG070 Room R11 STATISTICS FOR HILBERT SPACES II Chair: Gil Gonzalez-Rodriguez

E0294: **Simultaneous confidence bands for functional parameters**
Presenter: **Dominik Liebl**, University Bonn, Germany
Co-authors: Matthew Reimherr

Quantifying uncertainty using confidence regions is a central goal of statistical inference. Despite this, methodologies for confidence bands in Functional Data Analysis are underdeveloped compared to estimation and hypothesis testing. A major leap forward in this area is made by presenting a new methodology for constructing simultaneous confidence bands for functional parameter estimates. These bands possess several striking qualities: (1) they have a nearly closed-form expression, (2) they give nearly exact coverage, (3) they have a finite sample correction, (4) they do not require an estimate of the full covariance of the parameter estimate, and (5) they can be constructed adaptively according to the desired criteria. One option for choosing bands we find especially interesting is the concept of fair bands which allows us to do fair (or equitable) inference over subintervals. It could be especially useful in longitudinal studies over long time scales. Our bands are constructed by integrating and extending tools from Random Field Theory, an area that has yet to overlap with Functional Data Analysis.

E0937: **A goodness-of-fit test for the functional linear model with functional response**
Presenter: **Javier Alvarez-Liebana**, University of Oviedo, Spain
Co-authors: Wenceslao Gonzalez-Manteiga, Eduardo Garcia-Portugues, Gonzalo Alvarez-Perez

Functional data analysis enables to exploit the complexity and richness of data measured over continuous domains. When two functional random variables are available, it may be useful to determine their relation using a regression model. If the regression function is a linear Hilbert-Schmidt operator between two L_2 spaces, we are under the functional linear model with a functional response. We propose a novel goodness-of-fit test for the null (composite) hypothesis of this model, against a general, unspecified alternative. The test statistic is formulated in terms of the quadratic norm over a doubly-projected empirical process. It is easy to compute, interpret and calibrate on its distribution via a wild bootstrap on the residuals. A flexible hybrid approach, involving LASSO regularization and linearly-constrained least-squares, is used to perform the selection of functional predictors when estimating the residuals. The finite sample behavior of the test, regarding power and size, is illustrated via a complete simulation study under varying scenarios. The test is applied to some real datasets to check the validity of the model.

E0909: **A two-steps specification test for functional time series**
Presenter: **Alejandra Lopez-Perez**, Universidade de Santiago de Compostela, Spain
Co-authors: Javier Alvarez-Liebana, Wenceslao Gonzalez-Manteiga, Manuel Febrero-Bande

The functional data analysis framework allows the representation of continuous-time stochastic processes as sequences of random variables in function spaces. Focusing on Hilbert spaces, the autoregressive Hilbertian process (ARH) plays a central role in modeling time-series dynamics. We propose a two-steps test for the null composite hypothesis of the autoregressive Hilbertian model for a given order z, ARH(z), against a general alternative. The approach is twofold: we check if the functional sample and its lagged functional values are related via a Functional Linear Model with Functional Response (FLMFR) and, through the construction of linear representations, we propose a specification test for stochastic diffusion models. The later is a two-stage methodology where we first check the null hypothesis of the FLMFR and, secondly, under linearity, a functional F-test is performed. As an example, the Ornstein-Uhlenbeck process is characterized as ARH(1) to illustrate the finite sample performance of the proposed test. The new methodology is also applied to real datasets.

E0891: **On the complexity of functional data: A small ball probability approach**
Presenter: **Enea Bongiorno**, Universita del Piemonte Orientale, Italy
Co-authors: Aldo Goia, Philippe Vieu

The aim is to illustrate some recent developments concerning the Small-Ball Probability (SmBP) of a Hilbert-valued process. In particular, it is assumed that the SmBP can be factorized in two terms that play the role of a surrogate density and of a volumetric term. The latter factor is considered because it carries information about the complexity of the underlying process. An empirical estimator and some asymptotics are presented. Some applications are illustrated, in particular in estimating some complexity measures.

E0473: **Spatial multiscale analysis of functional count models**
Presenter: **Antoni Torres**, University of Malaga, Spain
Co-authors: Maria Pilar Frias Bustamante, Jorge Mateau, Maria Dolores Ruiz-Medina

A heterogeneity analysis is performed through different scales to describe the local variability of the sample paths of an infinite-dimensional spatial intensity in the context of a log-Gaussian Cox count model. A more flexible Functional Data Analysis (FDA) preprocessing procedure is implemented, allowing a higher degree of local singularity, avoiding over-smoothing. The derived multiresolution approximation reflects the space-time interaction at different scales, affecting the evolution of the log-intensity process that governs the counts. A wavelet-based nonparametric framework is adopted for spatial functional prediction. The approach presented is implemented for spatiotemporal prediction of respiratory disease

mortality at Spanish provinces.

EP002 Room Poster 1 POSTER SESSION I **Chair: Elena Fernandez Iglesias**

E0252: **Distributed fixed-point smoothing estimators in sensor networks connected by a directed graph**
Presenter: **Josefa Linares-Perez**, Universidad de Granada, Spain
Co-authors: Raquel Caballero-Aguila, Aurora Hermoso-Carazo

The focus is on the distributed fixed-point smoothing problem of discrete-time stochastic signals from measurements provided by a sensor network with a topology represented by a directed graph. The signal evolution model is assumed to be unknown and only the mean and covariance functions of the processes involved in the sensor measurement equations are available. The sensor measurements present random failures modelled by random parameter matrices and additive noises, which are assumed to be cross-correlated at the same time and correlated with the signal at the same and subsequent time steps (situation that occurs, for example, in state-space models when the sensor noises at each instant are correlated with the system noise at the previous time). The distributed estimation is performed in two stages; in the first one, every sensor node collects its own measurements and also, according to the network topology, those from the sensors within its neighbourhood in order to generate, through an innovation approach, recursive intermediate least-squares linear fixed-point smoothing estimators. In the second stage, the proposed distributed fixed-point smoothing estimators are generated in each sensor node as the optimal matrix-weighted linear combination of the intermediate estimators provided by its neighbours, according to the mean squared error criterion.

E0255: **Centralized estimation from measurements randomly subject to deception attacks**
Presenter: **Raquel Caballero-Aguila**, Universidad de Jaen, Spain
Co-authors: Aurora Hermoso-Carazo, Josefa Linares-Perez

Cyber-attacks are becoming one of the most popular deliberate ways to reduce the reliability of a network. A typical kind of cyber-attacks is the so-called deception attack, which may include a wrong sensor measurement or control input, an incorrect time-stamp or a wrong identity of the sending device. The focus is on the least-squares linear centralized estimation problem for discrete-time stochastic signals from measured outputs provided by different sensors, which transmit their observations to a processing center where the estimator is designed. Deception attacks to the sensor network are assumed to be launched by an adversary and the success probabilities of these attacks, which may be different for each sensor, are known. The false information sent by the adversary involves two components: one that neutralizes the actual measurements and a noise component, which is the added blurred information. Hence, at each sampling time, the processing center may receive from each sensor either the actual measurement or just the noise injected by the adversary. Using the information received and by an innovation approach, a recursive centralized linear filtering algorithm is obtained without requiring full knowledge of the signal evolution model, but only the first- and second-order moments of the processes involved in the observation equations.

E0260: **Fusion estimation in networked systems using fading measurements subject to transmission delays and losses**
Presenter: **Aurora Hermoso-Carazo**, Universidad de Granada, Spain
Co-authors: Raquel Caballero-Aguila, Josefa Linares-Perez

The signal estimation problem in multisensor systems has gradually become a meaningful topic of research in recent years. It has been well recognized that random imperfections are frequently found in networked systems which, if not addressed properly, are likely to impair the estimators performance. For this reason, considerable effort has been directed towards the analysis of models involving these phenomena and the design of estimation algorithms that do not neglect their effects. One of the most common phenomena of uncertainty in networked systems is the fading or degradation of the measurements, that can be due, for example, to restrictions of the physical equipment, or inaccuracy of the measurement devices. In addition, during the transmission to the processing center, the data packets may suffer random delays and/or losses, owing to unreliable communications, limited-capability or congestions. The focus is on the fusion estimation problem in networked systems from fading measurements, by assuming that the transmission is subject to random one-step delays and non-consecutive losses, and these uncertainties occur with different rates at the different sensors. Using a covariance-based approach and compensating the losses with the last measurement received, a recursive algorithm is designed for the distributed fusion estimation problem.

E0617: **Exact inference for an exponential distribution under generalized adaptive progressive hybrid censoring scheme**
Presenter: **Kyeongjun Lee**, Daegu University, Korea, South
Co-authors: Seonghee Park, YeongEun Hwang

In reliability studies and life-testing experiments, the failure time data of experimental items are often not completely available. This case is called to as censoring, and a type I and II censoring schemes are typical censoring scheme. We propose a new type censoring scheme named a generalized adaptive progressive hybrid censoring scheme. This censoring scheme is designed to correct the drawbacks in the adaptive progressive hybrid censoring scheme. Furthermore, we discuss inference for one-parameter exponential distribution under generalized adaptive progressive hybrid censoring scheme. We derive the moment generating function of the maximum likelihood estimator of the scale parameter of exponential distribution and the resulting confidence interval under generalized adaptive progressive hybrid censoring scheme.

E0712: **Density estimation for censored and contaminated data**
Presenter: **Elif Akca**, KU Leuven, Belgium
Co-authors: Ingrid Van Keilegom

A vast literature exists on covariate measurement error correction in a survival context, i.e., a variety of methods are available when an uncontaminated survival outcome is regressed on error-prone covariates. However, it is also possible that the measurements for the survival outcome are not error-free. When those measurements are censored, both censoring and measurement error should be taken into account. A flexible approach for density estimation in the presence of censoring and measurement error is proposed when no auxiliary variable or validation data are available. A classical additive measurement error model with Gaussian noise and a right-censoring scheme is assumed. It is shown that the assumed model is identifiable under certain conditions on the support of the censored-contaminated survival outcome and a methodology using Laguerre polynomials is offered for density estimation. The numerical performance of the proposed methodology is investigated on both simulated and real data.

E1197: **Observed trends in daily data and maximum hourly intensity of rainfall events in Asturias (NW Spain)**
Presenter: **Ana Belen Ramos-Guajardo**, University of Oviedo, Spain
Co-authors: Elena Fernandez Iglesias, Gil Gonzalez-Rodriguez

Extreme rainstorms in headwater catchments imply flash floods and landslides, the most serious geomorphological hazards in Asturias mountain region, located in the Cantabrian Range (NW of Spain). Brief episodes (less than 24 hours) of heavy rainfall are recognized as one of the most important triggers of these natural hazards, but the studies developed until now only include daily precipitation data. Due to climate change, the frequency and magnitude of these events are being altered. The main goal is to analyse long-term trends selecting hourly precipitation data from meteorological gauges. We have identified rainfall storm events by considering a 0.9995 quantile in the hourly precipitation series. For each rain gauge, a Bootstrap isotonic test in the strict sense has been applied to check whether the corresponding trend is constant or, otherwise, there is an effective increment. The obtained p-values at the usual significance levels have shown opposite behaviour between the inland and the coastal area. We identify an effective increment in expected total rainfall, total time and the number of storm events per year in the inland area. In contrast, an effective decrease is clear in the mean length and number of storm events in the coastal area. Although there is an important variability in Asturias,

the rainy trends in central/inner areas are more representative, where there is the highest activity of geomorphological hazards.

EP008 Room Poster 2 POSTER SESSION II **Chair: Elena Fernandez Iglesias**

E0384: **Estimation of log-odds ratio from group testing data using a Bayesian approach**
Presenter: **Viktor Skorniakov**, Vilnius University, Lithuania
Co-authors: Ugne Cizikoviene, Remigijus Leipus

Group testing (GT) is an important testing strategy with many applications spanning broad spectra of areas with a clinical setting (CS) among all the rest. In the case of CS, GT reduces to the testing of pooled specimens obtained by pooling untested individual specimen. A typical example occurs when, instead of testing k individual blood samples, one pools them, performs a single test, and finds out whether an infection is present in the pool. In case of a rare disease, the pool often tests negatively resulting in significant test cost savings. In case pool tests positively, retesting takes place if one wishes to identify all infected. This is called an identification task. However, if one wishes only to estimate disease prevalence (DP), subsequent retesting is not needed and leads to an estimation task (ET). There is a substantial body of literature indicating that GT is effective for ET when DP is low. We investigate the effectiveness of Bayesian GT based ET by conducting a pilot simulation study devoted to the case of two populations when the estimation of the odds ratio is a goal and test is imperfect.

E0597: **Confidence regions in a two-parameter model of congenital anomaly incidence in twins**
Presenter: **Jan Klaschka**, Institute of Computer Science of the Czech Academy of Sciences, Czech Republic
Co-authors: Jeno Reiczigel, Antonin Sipek

The focus is on m mutually independent pairs of binary variables with common expectation θ and equal correlation ϕ within each pair. The motivation and the main application field is the epidemiology of congenital anomalies (birth defects) in twins (variable coding: 1 = defect, 0 = no defect). Simultaneous confidence regions (CRs) are considered for θ and ϕ, based on numbers of pairs with two defects, and pairs with one defect. Fixed grid computational algorithms for CRs based on likelihood-ratio (LR) method, and Sterne (probability-based) method have been proposed and implemented in R. Both methods yield similar results in the sense that typically a CR of one type shares about 95% of its area with CR of the other type, none of the two CR areas is uniformly smaller, and the area difference is below 2%. The probability-based method guarantees, unlike the other method, coverage probability greater or equal to the nominal confidence level. On the other hand, advantages of the LR-based method over the Sterne method are a simpler and dramatically faster algorithm, as well as smooth CR boundaries (contrary to irregularly winding boundaries of Sterne CRs).

E0678: **Weighted average approach for joint modelling in absence of knowledge regarding the association structure**
Presenter: **Maha Alsefri**, University of Liverpool, United Kingdom
Co-authors: Ruwanthi Kolamunnage-Dona, Maria Sudell

Over the last decade, there has been an increasing interest in applying joint models to related longitudinal and time-to-event outcome data in clinical research, especially in studies with interest in examining patients repeatedly until an event of interest. The usage of this method is increasing due to their ability to account for informative dropout in the longitudinal data and to allow for the inclusion of time-varying covariates measured with error in the survival model. However, according to recent reviews of joint modelling, the current research is considerably limited in terms of specifying the association structure between the longitudinal and time-to-event outcomes in the absence of background knowledge. Information criteria (such as DIC) are generally applied to identify the best fit joint model from several potential association structures. We propose an alternative weighted averaging approach, which can be utilized to combine estimations from potential joint models. This results in parameter estimation being based on multiple different association structures instead of limiting to just one, possibly incorrect. Simulated data is used to investigate the proposed approach in both frequentist and Bayesian settings. Results of the simulation study, and real-world application from PBC (primary biliary cirrhosis) and HCC (Hepatocellular Carcinoma) studies, will be presented.

E0815: **On multi-factor state-space modelling and forecasting of EUA futures prices**
Presenter: **Jun Seok Han**, Macquarie University, Australia
Co-authors: Nino Kordzakhia, Pavel Shevchenko, Stefan Trueck

The European Union Emissions Trading System (EU ETS) was introduced in 2005 to confront rising greenhouse gas emissions. The EU ETS covers all major CO2 emitting industries, and next to European Emission Allowance (EUA) spot contracts, there is also a wide range of futures contracts available for trading. A multi-factor state-space model for risk-neutral pricing of EUA futures is presented that can also be applied for the out-of-sample forecasting of futures prices. A comparative analysis of the performance of a state-space model in a general setup with correlated measurement errors versus reduced-form models is conducted. As we deal with unobservable factors, we use the Kalman filtering technique for the estimation of the state variables, subsequently estimating the model parameters by maximising a marginal likelihood function. We illustrate the developed model, using a cross-section of daily futures contracts for the sample period from January 2013 - April 2020 that corresponds to the Phase III period of the EU ETS.

E1060: **Signal compression and Bayesian functional regressions in presence of hybrid principal component: An EEG-fMRI Dataset**
Presenter: **Mohammad Fayaz**, Shahid Beheshti University of Medical Sciences, Iran
Co-authors: Alireza Abadi, Soheila Khodakarim

In some situations that exist both scalar and functional data, called mixed and hybrid data, the hybrid PCA (HPCA) was introduced. Among the regression models for the hybrid data, we can count covariate-adjusted HPCA, the Semi-functional partial linear regression, FOF regression with signal compression, and functional additive regression, models. We study the effects of HPCA decomposition of hybrid data on the prediction accuracy of the two functional regression models: Bayesian scalar-on-function with Markov Chain Month Carlo (MCMC) sampler and function-on-function with signal compressions. We stated a two-step procedure for incorporating the HPCA in the functional regressions. The first step in reconstructing the data based on the HPCAs and the second step is merging data on the other dimensions and calculate the point-wise average of the desired functional dimension. We choose the number of HPCA based on the DIC, LPML, and deviance in the Bayesian settings and MSE and MPE for both of them. In the three simulations, we show that the regression models with the first HPCA have the best accuracy prediction and model fit summaries among no HPCA and all HPCAs with training/testing approach. Finally, we applied our methodology to the EEG-fMRI dataset.

CO097 Room R04 REALIZED MEASURE ANALYSIS AND MODELING IN HIGH DIMENSION **Chair: Massimiliano Caporin**

C0180: **Detecting event-driven systemic cojumps and market crashes**
Presenter: **Deniz Erdemlioglu**, IESEG School of Management, France
Co-authors: Xiye Yang, Christopher Neely

A new testing framework is developed to detect the presence of systemic cojumps in a large panel of financial assets. The essential feature of the approach is that the test statistics are conditional on the release times of information events, which in turn allows us to pinpoint when precisely individual stocks or portfolio indices jump at the same time, or in a downward direction, when assets crash together systemically at high frequency. For inference, we introduce a computationally feasible StepM-type recursive bootstrapping procedure in high (asset and event) dimension, control for the multiple testing problem and eliminate spurious detection. We establish the bootstrap consistency of the tests and show in simulations that the tests have good size and considerably high power. Based on the high-frequency data on Dow Jones constituents and sector-specific ETFs, our

empirical analysis provides strong evidence of systemic cojumps driven by the FOMC news announcements and monetary policy shocks. We find that a large fraction of the detected systemic cojumps exhibit downward pattern, indicating that monetary policy actions can lead to sudden crashes and generate systemic downside risk. We discuss the practical implications of our results for portfolio diversification, (news-driven) systemic risk monitoring and realized stress testing.

C0186: **Jumps in realized volatility modeling and forecasting: Empirical evidence and a new model**
Presenter: **Massimiliano Caporin**, University of Padova, Italy

Building on an extensive empirical analysis, the relevance of jumps and signed variations in predicting Realized Volatility is investigated. We show that properly accounting for intra-day volatility patterns and staleness sensibly reduces the identified jumps. Realized Variance decompositions based on intra-day return size and sign improve the in-sample fit of the models commonly adopted in empirical studies. We also introduce a novel specification based on a more informative decomposition of Realized Volatility, which offer improvements over standard models. From a forecasting perspective, the empirical evidence shows that most models, irrespective of their flexibility, are statistically equivalent in many cases. This result is confirmed with different samples, liquidity levels, forecast horizons and possible transformations of the dependent and explanatory variables.

C0358: **Price jumps and cross-company news impact**
Presenter: **Francesco Poli**, University of Padova, Italy
Co-authors: Massimiliano Caporin

Evidence is provided about how macroeconomic scheduled news, firm-specific unscheduled news and liquidity drive real-time jump spillovers among stocks belonging to different economic sectors. To this end, we employ high-frequency data of 220 constituents of the Russell 3000 index equally divided into eleven sectors. We use multiple logistic regression with interactions and penalization methods to distinguish the different roles played by the surprise component of the macroeconomic news and the sentiment of the firm-specific news.

C0614: **Realized moments: Identification and pricing**
Presenter: **Roberto Reno**, University of Verona, Italy
Co-authors: Davide Pirino, Federico Bandi, Aleksey Kolokolov

The purpose is to study the properties of realized high-order moments under a data generating process accounting for key stylized features: infrequent discontinuities in unobserved equilibrium prices and staleness in observed prices, a phenomenon linked to volume dynamics. The focus is on identification and pricing. In terms of identification, we show how the interplay between price discontinuities and prices staleness will, in general, lead to biased and/or noisy high- order moment estimates. We also show how a combination of thresholding and corrections for staleness-induced biases can be deployed to extract reliable information about high-order continuous and discontinuous variation. Regarding pricing, the use of thresholding and debiasing leads to ample evidence about the negative cross-sectional pricing of idiosyncratic price discontinuities at high frequency. We show that accounting for staleness is (1) important for the correct identification of high-order moments, (2) revealing about these moments cross-sectional pricing and (3) informative about the pricing of illiquidity, for which staleness is a rich proxy.

C0765: **Systematic staleness**
Presenter: **Davide Pirino**, University of Rome Tor Vergata, Italy
Co-authors: Federico Bandi, Roberto Reno

Asset prices are stale. We define a measure of systematic (market-wide) staleness as the percentage of small price adjustments over multiple assets. A notion of idiosyncratic (asset-specific) staleness is also established. For both systematic and idiosyncratic staleness, we provide a limit theory based on joint asymptotics relying on increasingly-frequent observations over a fixed time span and an increasing number of assets. Using systematic and idiosyncratic staleness as moment conditions, we introduce novel structural estimates of market liquidity and funding liquidity based on transaction prices only. The estimates yield revealing information about the dynamics of the two notions of liquidity and their interaction.

CO716 Room R08 BAYESIAN MACROECONOMETRICS Chair: Aubrey Poon

C0195: **Flexible mixture priors for time-varying parameter models**
Presenter: **Niko Hauzenberger**, University of Salzburg, Austria

Time-varying parameter (TVP) models often assume that the TVPs evolve according to a random walk. This assumption, however, might be questionable since it implies that coefficients change smoothly and in an unbounded manner. We relax this assumption by proposing a flexible law of motion for the TVPs in large-scale vector autoregressions (VARs). Instead of imposing a restrictive random walk evolution of the latent states, we carefully design hierarchical mixture priors on the coefficients in the state equation. These priors effectively allow for discriminating between periods where coefficients evolve according to a random walk and times where the TVPs are better characterized by a stationary stochastic process. Moreover, this approach is capable of introducing dynamic sparsity by pushing small parameter changes towards zero if necessary. The merits of the model are illustrated by means of two applications. Using synthetic data we show that our approach yields precise parameter estimates. When applied to US data, the model reveals interesting patterns of low-frequency dynamics in coefficients and forecasts well relative to a wide range of competing models.

C0287: **Horseshoe prior Bayesian quantile regression**
Presenter: **David Kohns**, Heriot-Watt University, United Kingdom
Co-authors: Tibor Szendrei

The Horseshoe Prior is extended to the Bayesian Quantile Regression (HS-BQR), a fast sampling algorithm is provided that speeds up computation significantly in high dimensions. The performance of the HS-BQR is tested on large scale Monte Carlo simulations and a high dimensional Growth-at-Risk (GaR) forecasting exercise for the U.S. The Monte Carlo design considers several sparsity structures (sparse, dense, block) and error structures (i.i.d. errors and heteroskedastic errors). Compared to alternative shrinkage priors, the proposed HS-BQR yields at worst similar, or better performance considered when evaluated using coefficient bias and forecast error. We find that the HS-BQR is particularly potent in sparse designs and when estimating extreme quantiles. The simulations also highlight that in order to identify quantile specific location and scale effects for individual regressors in dense DGPs, a lot of data are necessary. In the GaR application, we forecast tail risks as well as complete forecast densities using the McCracken database. Quantile specific and density calibration scoring functions show that the HS-BQR provides the best performance, especially at short and medium run horizons. The ability to produce well calibrated density forecasts and accurate downside risk measures in the face of large data contexts makes the HS-BQR a promising tool for nowcasting applications and recession modelling in the face of the Covid-19 pandemic.

C0745: **Macroeconomic uncertainty and the demand for secondary health insurance: Evidence from Bayesian quantile SVARs**
Presenter: **Annika Camehl**, Erasmus University Rotterdam, Netherlands
Co-authors: Kathrin Gruber

A Bayesian quantile structural vector autoregressive model is developed. It is applied to analyze the reaction of household secondary health insurance purchases to macroeconomic uncertainty shocks. We assume that the error terms follow a quantile-restricted overfitting finite Gaussian mixture distribution. This approach mitigates problems of employing the commonly used asymmetric Laplace distribution to the multivariate case. By analyzing quantiles, we uncover potential asymmetries in the reaction to uncertainty shocks from households with purchases above or below

the median. We take variation in secondary health insurance spending, based on U.S. household expenditure data, as an indication that households have different risk perceptions. We then trace back heterogeneities in households' reactions to their socioeconomic characteristics. Our preliminary findings illustrate that household spending on secondary insurances products increases in response to uncertainty shocks. We identify these shocks by using exogenous proxy information. The intensity of the responses varies across quantiles, especially for households in birth cohorts in the early thirties and those in retirement age, with the strongest reaction visible for higher quantiles.

C0385: **Monthly GDP estimates for the US states**
Presenter: **Aristeidis Raftapostolos**, University of Strathclyde, United Kingdom
Co-authors: Gary Koop, Stuart McIntyre, James Mitchell

Models are developed for regional nowcasting by producing monthly nowcasts and historical estimates of GDP growth at the US state level using a Mixed frequency Vector Autoregression (MF-VAR). MF-VARs have enjoyed great popularity in policy circles since they can provide timely, high frequency nowcasts of low frequency variables such as GDP growth which are released with a delay. A common set-up is to nowcast a quarterly variable (e.g. GDP growth) using several monthly variables. Nowcasting state level GDP growth is more of a challenge since there are 51 (50 states plus District of Columbia) variables to be nowcast and the frequency mismatch is more complicated (i.e. we have a three-way frequency mismatch involving annual, quarterly and monthly variables) and changes over time. We work with MF-VARs which are much larger than is conventional, involve many more, and more complicated, latent states. This raises challenges in terms of over-parameterization concerns and the computational burden. We develop a Bayesian modelling framework which overcomes these challenges and present results on the accuracy of nowcasts of real-time economic growth in the USA from 2006 to 2019 at the monthly frequency.

C0271: **Reconciled estimates of monthly GDP in the US**
Presenter: **Aubrey Poon**, University of Strathclyde, United Kingdom
Co-authors: Gary Koop, Stuart McIntyre, James Mitchell

In the US, income and expenditure side estimates of GDP (GDPI and GDPE) measure true GDP with error and are available at the quarterly frequency. Methods exist for producing reconciled quarterly estimates of GDP based on GDPI and GDPE. We extend these methods to provide reconciled historical GDP estimates at the monthly frequency from 1960. We do this using a Bayesian Mixed Frequency Vector Autoregression involving GDPE, GDPI, unobserved true GDP and monthly indicators of short-term economic activity. We illustrate how the new monthly data contribute to our historical understanding of business cycles.

CC807 Room R02 CONTRIBUTIONS IN TIME SERIES ECONOMETRICS Chair: Maria Kyriacou

C0897: **Variable selection for fractional time series with shifts in the deterministic component**
Presenter: **Mustafa Kilinc**, WHU - Otto Beisheim School of Management, Germany

Shifts in the deterministic component of time series occur frequently. Traditional estimation methods ignore this feature of time series. An easy and straightforward solution is to insert indicators to the regression model. However, often ex-ante, there is no knowledge about the number of shifts and the locations of the shifts. A saturation technique has been proposed that inserts all the possible indicators into the regression model, so one for each observation, and removes the irrelevant ones using a simple t-test. This approach is extended in two main directions. First, we consider a parametric error component composed of fractional time series and estimate the model using a GLS-type of estimator. The memory parameter, which characterizes the behaviour of the stochastic component of the model, is considered unknown but lying in an arbitrarily large interval. Second, the deterministic part is an additive generalized polynomial trend with a known exponent parameter. Special cases that arise in this setting are level shifts and trend shifts. Monte Carlo simulations, confirm the accuracy of nominal significance levels under null and the precision of finding the shifts under the alternative. An empirical application is considered for the monthly core consumer price inflation in the US and UK.

C0996: **Estimation of ARMA models with t-distributed innovations**
Presenter: **Haruhisa Nishino**, Hiroshima University, Japan

The standard ARMA model assumes that its innovations are white noise processes with 0 mean and a constant variance. That is, the ARMA model is a second-order stationary process characterized by its autocovariance function. The white noise process has no information about its fourth-order moment. In general, assuming a Gaussian process and a Gaussian likelihood enables us to estimate the ARMA model. On the other hand, the literature of financial time series tells us that the financial returns have fatter tails than Gaussian ones. The student t-distribution is a typical example of fat-tailed distributions. The fat-tailed property is related to the fourth-order moment. We thus consider that estimation for ARMA models with t-distributed innovations, which is useful for analyzing financial time series. If we know degrees of freedom of the t-distribution of the model, is it possible to estimate the parameters of the ARMA model more efficiently than the Gaussian likelihood? Also, the talk proposes a preliminary estimate for degrees of freedom based on the method of moments, since estimating degrees of freedom of t-distribution by MLE causes a severe problem.

C0446: **A simple unit root test consistent against any stationary alternative**
Presenter: **Frederique Bec**, THEMA University of Cergy-Pontoise and CREST, France
Co-authors: Alain Guay

A $t-$like unit root test is proposed, which is consistent against any stationary alternatives, nonlinear or noncausal ones included. It departs from existing tests in that it uses an unbounded, not adaptive set of thresholds. In our setup, thanks to the straightforward nonlinear stationary alternative specification and the particular choice of the thresholds set, the proposed unit root test contains the standard ADF test as a special case. This, in turn, yields a sufficient condition for consistency against any ergodic stationary alternative. From a Monte-Carlo study, it turns out that the power of our unbounded non-adaptive tests, in their average and exponential versions, outperforms existing bounded tests, either adaptive or not. This is illustrated by an application to interest rate spread data.

C0450: **Multivariate Wold decompositions**
Presenter: **Federico Severino**, Universite' Laval, Canada
Co-authors: Fulvio Ortu, Claudio Tebaldi, Simone Cerreia-Vioglio

Multivariate time series are driven by a collection of (possibly correlated) univariate shocks that need to be properly identified in the applications. Moreover, in many economic contexts, the process under scrutiny is the outcome of the superposition of simultaneous disturbances with heterogeneous frequencies that can generate short-, medium- or long-term effects. Given a weakly stationary vector process, we provide a methodology to elicit uncorrelated persistent components driven by multivariate shocks with increasing duration: the Multivariate Extended Wold Decomposition. By introducing multivariate scale-specific responses, we can quantify the persistence in vector autoregressive models, once their shocks are identified. To derive the decomposition, we embed the vector process in a Hilbert A-module framework where matrices replace the field of scalars, and we prove the Abstract Wold Theorem for self-dual pre-Hilbert A-modules. From this abstract result, by using projection techniques, we retrieve the well-known Multivariate Classical Wold Decomposition. We also derive the persistence-based Multivariate Extended Wold Decomposition. We finally apply the latter to some well-known macroeconomic bivariate VAR models.

C1050: **The fractional sinusoidal waveform process**
Presenter: **Federico Maddanu**, University of Rome Tor Vergata, Italy

Co-authors: Tommaso Proietti

A novel model for time series displaying persistent cycles, the fractional sinusoidal waveform process, is proposed. It is based on the simple idea of allowing the parameters that regulate the amplitude and the phase of a cycle to evolve according to a fractional noise process. While the autoregressive polynomial of the reduced form is a Gegenbauer polynomial, the main advantage of our formulation is that the autocovariance function is available in closed form. This opens the way for estimation of the parameters by exact maximum likelihood, evaluated with the support of the Durbin-Levinson algorithm.

CG026 Room R03 CONTRIBUTIONS IN FINANCIAL ECONOMETRICS III Chair: Sandra Paterlini

C0557: **A free-knot spline-GARCH model**

Presenter: **Oliver Old**, FernUniversitat in Hagen, Germany

Global estimation of parameters in GARCH models could easily lead to the premature conclusion of a nearly integrated volatility process due to a very strong volatility persistence. This could be caused by the erroneous assumption of a constant unconditional variance over the entire sample, in particular for long financial time-series. The assumption of constant unconditional variance is taken into account by volatility models with a multiplicative decomposition of the conditional variance into a short-term and a long-term component. The short-term component is represented by an asymmetric GJR-GARCH model, and the time-varying long-term volatility is modelled by a B-spline function. The location of the knots affects the shape of the spline-function. The main contribution is a free-knot spline smoothing approach. Therefore, knot locations are not given in advance but rather estimated within the optimisation routine with all other parameters. Besides, enhanced mitigation of the volatility persistence, the aim is to test whether forecast accuracy is improved compared to the spline-GARCH model. That would imply a good approximation of the spline-function to the data and a great improvement for modelling time-varying unconditional variance. Therefore, the free-knot spline GARCH model is investigated by a comprehensive simulation study and by the S&P500 composite index.

C0674: **Conditional asymmetry in ARCH(∞) models**

Presenter: **Julien Royer**, CREST, France

An extension of ARCH(∞) models is considered to account for conditional asymmetry in the presence of high persistence. After stating existence and stationarity conditions, we develop the statistical inference of such models and proves the consistency and asymptotic distribution of a Quasi Maximum Likelihood estimator. Some particular specifications are studied, and tests for asymmetry and GARCH validity are derived. Finally, we present an application on a set of equity indices to reexamine the preeminence of GARCH-type specifications. We find strong evidence that the short memory feature of such models is not suitable for lightly traded assets.

C0445: **Ex-ante industry-based uncertainty network and the business cycle**

Presenter: **Mattia Bevilacqua**, London School of Economics, United Kingdom

Co-authors: Jozef Barunik, Robert Faff

A forward-looking measure of uncertainty network connectedness for the US industries is developed through a time-varying parameter VAR (TVP VAR) model. Our measure is constructed from options investors' expectations about the next month uncertainty of the US industries. We rank the dynamics of each industry uncertainty based on the contribution to the whole system and in relation to the business cycle in a dynamic way. We uncover a predominant role for communications, industrials and information technology industries in terms of uncertainty propagation mechanism throughout the cycles, being these denoted as "uncertainty hubs". Other industries such as materials, real estate and utilities are classifiable as "uncertainty not-hubs". We detect the ex-ante network of uncertainty as a useful predictor of business cycles and macroeconomic indicators, showing even greater predictive ability when extracted from uncertainty hubs only.

C0183: **How to measure oil market uncertainty: An application of Google Trends**

Presenter: **Esti Tri Widyastuti**, University of Aberdeen, United Kingdom

Co-authors: Marc Gronwald

An oil market specific uncertainty measure is proposed based on Google Trends. Frequent Google searches for terms such as oil prices, OPEC, and oil demand are assumed to capture oil market uncertainty. A careful comparison of this newly proposed uncertainty measure with some recently proposed in the literature shows a remarkable degree of similarity. In other words, the crude oil market is found to be an important source of both economic and financial market uncertainty. Furthermore, the empirical relationship between the newly proposed measure and some oil market specific variables such as oil exploration activity is analysed using a standard VAR approach. The results indicate that uncertainty shocks affect exploration and, in turn, future oil production, negatively and significantly.

CG256 Room R06 CONTRIBUTIONS IN MODELLING, VOLATILITY AND ACCURACY Chair: Toshiaki Watanabe

C0764: **Model risk of volatility models**

Presenter: **Emese Lazar**, University of Reading, United Kingdom

Co-authors: Ning Zhang

To evaluate the accuracy of volatility models, we propose a new model risk measure and estimation methodology based on loss functions. The reliability of the proposed estimation has been verified via simulations, and the estimates provide a reasonable fit to the true model risk measure. We undertake an empirical analysis based on several assets, identify the models most affected by model risk, and argue that the accuracy of volatility models can be improved by adjusting variance forecasts for model risk. We find that after crises, the model risk increases especially for badly fitting volatility models.

C0803: **CoVaR with volatility clustering, heavy tails and non-linear dependence**

Presenter: **Giorgia Rivieccio**, Parthenope University, Italy

Co-authors: Giovanni De Luca, Michele Leonardo Bianchi

The conditional value-at-risk is estimated by fitting different multivariate parametric models capturing some stylized facts about multivariate financial time series of equity returns: heavy tails, negative skew, asymmetric dependence, and volatility clustering. While the volatility clustering effect is got by AR-GARCH dynamics of the GJR type, the other stylized facts are captured through non-Gaussian multivariate models and copula functions. The CoVaR is computed on the basis on the multivariate normal model, the multivariate normal tempered stable (MNTS) model, the multivariate generalized hyperbolic model (MGH) and four possible copula functions. These risk measure estimates are compared to the CoVaR based on the multivariate normal GARCH model. The comparison is conducted by backtesting the competitor models over the time span from January 2007 to March 2020. In the empirical study we consider a sample of listed banks of the euro area belonging to the main or to the additional global systemically important banks (GSIBs) assessment sample.

C0185: **Modelling price and volatility jump clustering by marked Hawkes processes**

Presenter: **Jian Chen**, University of Reading, United Kingdom

Co-authors: Michael Clements, Andrew Urquhart

Clustering behaviours of price and volatility jumps are studied using high-frequency data, modelled using a Marked Hawkes Process embedded in a bivariate jump-diffusion model. Under de-periodisation, we find evidence showing self-excitation behaviours of jumps in both individual stocks and an index. Also, considering positive, negative price jumps and volatility jumps, the impact that an occurrence of a jump in one dimension has

on that in another dimension is shown to be asymmetry. More importantly, the extent of this impact is shown empirically to be positively correlated with jump size. We also formalise the self-excitement and self-freeze properties of durations between two jumps. More self-freeze behaviours have been found in empirical studies. We estimate model parameters using Bayesian inference by Markov Chains Monte Carlo.

C0251: **Connectedness between the crude oil and equity markets during the pre-and post-financialisation era**
Presenter: **Sania Wadud**, The University of Aberdeen, United Kingdom
Co-authors: Robert B Durand, Marc Gronwald

The financialisation of commodities may change the nature of price volatility and connectedness between equity and commodity futures market. We investigate whether the link between equity and crude oil futures markets in their return volatility has been altered by financialisation by accounting for the systematic patterns of commodity price volatility, namely, seasonality and maturity effects for the period 1993-2019; using weekly data partitioned into pre-and-post financialisation periods. We adopt VAR-DCC-GARCH model to estimate conditional volatility and time-varying correlation to assess how their dynamics have evolved. The conditional volatility and the conditional correlation of crude oil and equities are found to be positively linked after financialisation period. We find that speculation (open interest) has a negative impact on crude oil futures price volatility before (during) financialisation period. Additionally, we use Granger causality analysis to inspect the existence of lead-lag relations among price volatility, correlation, speculation, and open interest and find that speculative activity leads to conditional volatility during financialisation period. The estimated results support the Samuelson hypotheses for both sample periods; however, this effect is found to be diminishing in the financialisation period.

C0977: **Modelling volatility cycles: The $(MF)^2$ GARCH model**
Presenter: **Christian Conrad**, Heidelberg University, Germany
Co-authors: Robert Engle

A multiplicative factor multi-frequency ($(MF)^2$) component GARCH model is proposed. The model consists of a short-term GARCH component and one or multiple long-term components. The long-term components are based on MEM equations for the average standardized forecast errors of the GARCH component and capture the counter-cyclical behavior of financial volatility. We derive conditions weak stationarity of the $(MF)^2$ GARCH and discuss the news impact function. Since the new model is dynamically complete, it is straightforward to construct multi-step ahead volatility forecasts. We apply the model to forecast the volatility of the S&P 500 and three international stock markets. We show that the long-term component of the S&P 500 behaves counter-cyclical and is driven by news about the macroeconomic outlook. The $(MF)^2$ GARCH significantly outperforms the nested one-component GJR GARCH in out-of-sample forecasting.

CG024 Room R07 CONTRIBUTIONS IN CREDIT RISKS Chair: Jonathan Crook

C0193: **Recovery process optimization using survival regression**
Presenter: **Jiri Witzany**, University of Economics in Prague, Czech Republic
Co-authors: Anastasiia Kozina

The goal is to propose, empirically test and compare different logistic and survival analysis techniques in order to optimize the debt collection process. This process uses various actions, such as phone calls, mails, visits, or legal steps to recover past due loans. We focus on the soft collection part, where the question is whether and when to call a past-due debtor with regard to the expected financial return of such an action. We propose using the survival analysis technique, in which the phone call can be compared to a medical treatment, and repayment to the recovery of a patient. We show on a real banking dataset that, unlike ordinary logistic regression, this model provides the expected results and can be efficiently used to optimize the soft collection process.

C0327: **Credit rating downgrade risk on equity returns**
Presenter: **Periklis Brakatsoulas**, Charles University, Faculty of Social Sciences, Czech Republic
Co-authors: Jiri Kukacka

A four-factor model is developed which intends to capture size, value, and credit rating transition patterns in excess returns for a panel of predominantly mid- and large-cap entities. Using credit transition matrices and rating histories from 48 US issuers, we provide evidence to support a statistically significant negative downgrade risk premium in excess returns, suggesting that stocks at higher risk of failure tend to deliver lower returns. The performance of the model remains robust across several estimation methods. Panel Granger causality test results indicate that there indeed is a Granger-causal relationship from credit rating transition probabilities to excess returns. Our paper thus provides a new methodology to generate firm-level downgrade probabilities and the basis for further empirical validation and development of Fama-French-type models under financial distress.

C0492: **Stress testing behavioural and macroeconomic risks for credit portfolios**
Presenter: **Jonathan Crook**, University of Edinburgh, United Kingdom
Co-authors: Viani Djeundje

Large banks are required to stress test their credit portfolios annually under Basel II. Stress testing credit portfolios to macroeconomic shocks at account level involves parameterising a model predicting the probability of default followed by hypothesising specific shocks or by simulation to derive a value at risk (VaR) or expected shortfall (ES) 12 months into the future. The simulation requires that the simulated values of the macroeconomic variables are mutually consistent. But the probability of default is also correlated with time-varying behavioural variables, which in turn are correlated with the macroeconomy. Simulation studies have estimated the VaR when mutually consistent macroeconomic values have been simulated or when behavioural variables have been simulated but not when both are simulated. We present a method to simulate both behavioural and macroeconomic variables 12 months into the future whilst maintaining the correlation structure between them to derive a more comprehensive simulation methodology to stress test a credit portfolio.

C1030: **Copula-Heckit: Application of modified models with selectivity for loss amount prediction in case of bank failures**
Presenter: **Henry Penikas**, Higher School of Economics, Russia

A range of papers deals with the probability of bank default modeling. A separate stack of papers discusses bank failure case studies. However, some works do not limit themselves to the probability of failure prediction. They attempt to predict the loss amount in case of a bank failure. The Heckit model is used. It presumes a correlation of errors from the selection equation and the principal one. However, such a Pearson correlation coefficient does not capture the rich bivariate dependence patterns that might occur in the real-world. Copulas, including Archimedean ones, do allow for this. The objective is to demonstrate the advantage of modifying Heckit model to account for copulated errors, i.e. to use copula-Heckit model. We start with the simulated data and proceed with the empirical on. It is Russian bank license withdrawal data for 2013-2020. We show how copula-Heckit model outperforms the conventional one.

C1116: **Measuring the default risk of small business loans: Improving credit risk prediction using deep learning**
Presenter: **Yiannis Dendramis**, Athens University of Economics and Business, Greece
Co-authors: Elias Tzavalis, Aikaterini Cheimarioti

A multilayer artificial neural network (ANN) method, known as deep learning ANN, is suggested to predict the probability of default (PD) within the survival analysis framework. Deep learning ANN structures consider hidden interconnections among the covariates determining the PD, which

can lead to prediction gains compared to parametric statistical methods. The application of the ANN method to a large data set of small business loans demonstrates prediction gains for the method relative to the logit and skewed logit models. These gains mainly concern short term prediction horizons. They are more apparent for the type I misclassification error of loan default events, which has important implications for bank loans portfolio management. To identify the effects of covariates on the PD by the ANN structure, the paper proposes a bootstrap sampling method obtaining the distribution of changes of the PD over discrete covariate changes, while controlling for possible interactions among the covariates. We find that the covariates with the most important influence on the PD include the delinquent amount of a loan over its total balance, the payments and the balance of the loan over its instalment, as well as the delinquency buckets of a loan. The duration of a loan is also found to be an important factor of default risk.

Monday 21.12.2020 14:15 - 15:55 Parallel Session P – CFE-CMStatistics

EI009 Room R11 ALGORITHMS AND HIGH STAKES POLICY DECISIONS Chair: Sofia Olhede

E0165: **Stop making excuses for black-box models**
Presenter: **Cynthia Rudin**, Duke University, United States

With the widespread use of machine learning, there have been serious societal consequences from using black-box models for high-stakes decisions, including flawed bail and parole decisions in criminal justice. Explanations for black-box models are not reliable and can be misleading. If we use interpretable machine learning models, they come with their own explanations, which are faithful to what the model actually computes. Several reasons will be given why we should use interpretable models, the most compelling of which is that for high stakes decisions, interpretable models do not seem to lose accuracy over black boxes - in fact, the opposite is true, where when we understand what the models are doing, we can troubleshoot them to gain accuracy ultimately.

E0168: **Politics, infrastructures and design choices in the DP-3T contact tracing protocol**
Presenter: **Michael Veale**, University College London, United Kingdom

During the COVID-19 pandemic, policy-makers looked eagerly at mobile apps to avoid lockdown or re-open society. As a scientific intervention, we formed an international consortium to create an open protocol and codebase called Decentralised Privacy-Preserving Proximity Tracing (DP-3T). It enables smartphone owners to be notified of a significant contact event with a later diagnosed individual without requiring a centralised database or persistent identifiers. We will describe our design choices and motivations, including privacy and purpose limitation by design and graceful degradation. We will reflect upon the role of the large technology platforms, particularly Apple and Google, in arbitrating between choices and designs. Their choices ultimately led to the adoption of the DP-3T standard around the world, and as scholars, we must closely reflect on their power to create specific computational and statistical infrastructures. We will also discuss desires to create population-scale statistics in a privacy-preserving manner. We will argue a narrow focus on privacy in these contexts, potentially able to deliver serious societal harm.

E0181: **Public confidence in statistical models: The UK approach to exam grades in summer 2020**
Presenter: **Ed Humpherson**, Office for Statistics Regulation, United Kingdom

The Office for Statistics Regulation (OSR), the official statistics regulator in the UK, has been undertaking a review of the use of statistical models to award exam grades to school children and other learners in the UK in the summer of 2020. The policy intention was that grades would be awarded using statistical models that brought together prior attainment, teacher judgement and past performance of the school they attended. But when the grades were released, there was a widespread public concern, and the modelled grades were abandoned. The aim is to explain how the OSR review, to be published in early 2020, will focus on public confidence and how public confidence needs to be interwoven throughout the process of model development.

EO562 Room R12 FUNCTIONAL DATA ANALYSIS Chair: Marie-Helene Descary

E0368: **Conditional independence testing for functional data**
Presenter: **Anton Rask Lundborg**, University of Cambridge, United Kingdom
Co-authors: Rajen D Shah, Jonas Peters

The aim is to study the problem of testing the null hypothesis that X and Y are conditionally independent given Z, where each of X, Y and Z may be functional random elements. We show that even in the idealised setting where (X,Y,Z) are jointly Gaussian with Z infinite-dimensional (i.e., functional) any test with power β at an alternative, must reject some null with probability at least β. Given the untestability of this hypothesis, we argue that tests must be designed so their suitability for a particular problem may be judged easily. To this end, we propose regressing each of X and Y onto Z and then computing the Hilbert-Schmidt norm of the outer product of the resulting residuals. We show that the level of the resulting test is controlled uniformly over a class of distributions governed primarily by the requirement that the regressions estimate the conditional expectations of X and Y given Z sufficiently well. Whilst our result allows for arbitrary regression methods, we develop the theoretical guarantees for Tikhonov regularised regressions. Simulations studies demonstrate the effectiveness of our approach for conditional independence testing and its applications in variable selection and truncation point estimation in functional linear models.

E0377: **Sparsely observed functional time series: Estimation and regression**
Presenter: **Tomas Rubin**, EPFL, Switzerland
Co-authors: Victor Panaretos

Functional time series analysis has traditionally been carried out under the assumption of complete observation of the constituent series of curves, assumed stationary. Nevertheless, it may very well happen that the data available to the analyst are not the actual sequence of curves, but relatively few and noisy measurements per curve, potentially at different locations in each curve's domain. First, we construct the spectral-domain-based estimator of the latent functional time series dynamics from noisy samples. Second, the estimated dynamic correlations are used to predict latent curves by borrowing strength across time. And third, we extend the framework to the lagged regression model where one functional time series is regressed onto another and show how to perform estimation and prediction from sparse and noisy data. The methodology is illustrated by application to financial data set on the US Treasury yield curve, a sparsely observed functional time series, being regressed on a time series of macroeconomic variables.

E0553: **Functional experimental design via analytic permutation testing**
Presenter: **Adam Kashlak**, University of Alberta, Canada
Co-authors: Sergii Myroshnychenko, Susanna Spektor

Permutation testing is a powerful non-parametric testing procedure that applies to a variety of testing scenarios and requires few assumptions. The main downfall of the permutation test is the excessive computation time required to run such a test making it impractical in some settings and unusable in others. We rectify this via application of variants of the Kahane-Khintchine inequality to construct an analytic upper bound on the permutation test p-value. Our method applies to two-sample and k-sample testing for univariate, multivariate, and functional data. We demonstrate its usefulness on a dataset of spoken phonemes, which makes use of two experimental designs: the Latin square design and the randomized block design.

E1103: **Function-on-function mixture model clustering**
Presenter: **Susana Conde Llinares**, The University of Warwick, Alan Turing Institute, London, United Kingdom
Co-authors: Shahin Tavakoli, Daphne Ezer

Gene expression data is often collected over time in a variety of experiments under different experimental conditions. Genes may have very different temporal gene expression profiles, but adjusting their expression patterns in the same way through experimental conditions. We aim to find clusters that capture functional regression relationships between a temporal response and temporal explanatory variables, possibly more than one. We develop a K-means type iterative-consensus clustering algorithm in which each cluster is defined by a function-on-function regression model fitted using boosting. Our models allow for many situations, including even autoregressive random error terms inter alia. We validate them with extensive simulations and then apply them to identify groups of genes whose diurnal expression pattern is similarly perturbed by the season.

Our clusters are enriched for genes with similar biological functions, including one characterized with both photosynthesis-related functions and polysomal ribosomes.

EO433 Room R13 ADVANCES OF COMPLEX DATA ANALYSIS Chair: Xiaoke Zhang

E0527: **Broadcasted nonparametric tensor regression**
Presenter: **Raymond Ka Wai Wong**, Texas A&M University, United States
Co-authors: Ya Zhou, Kejun He

A novel broadcasting idea is proposed to model the nonlinearity in tensor regression non-parametrically. Unlike existing non-parametric tensor regression models, the resulting model strikes a good balance between flexibility and interpretability. A penalized estimation and corresponding algorithm are proposed. The theoretical investigation, which allows the dimensions of the tensor covariate to diverge, indicates that the proposed estimation enjoys desirable convergence rate. Numerical experiments are conducted to confirm the theoretical finding and show that the proposed model has an advantage over existing linear counterparts.

E0682: **Adaptive functional thresholding for sparse covariance matrix function estimation**
Presenter: **Qin Fang**, The London School of Economics and Political Science, United Kingdom
Co-authors: Xinghao Qiao

With the emergence of functional data in contemporary science and business, the problem of large covariance matrix function estimation arises in many applications. To consistently estimate covariance matrix function in high dimensions, we introduce a new class of functional thresholding operators, of which the thresholding and shrinkage conditions are imposed on function's Hilbert-Schmdit norm to encourage functional-sparsity, and propose an adaptive functional thresholding procedure of the sample covariance matrix function, taking into account the variability of functional entries. We further investigate the consistency and sparsistency of the proposed estimator. Monte Carlo simulations show the advantage of this estimator in terms of estimation accuracy and support recovery by comparing it with the universal functional thresholding estimator. As a motivating example, we study the functional connectivity using resting-state fMRI time-series data from two neuroscience datasets.

E0733: **Robust batch policy learning for indefinite-horizon Markov decision processes**
Presenter: **Zhengling Qi**, The George Washington University, United States

The indefinite-horizon Markov decision process (MDP) is considered where each policy is evaluated as a set of average rewards over different horizon lengths with different reference distributions. Given pre-collected data generated by some behavior policy, our goal is to learn a robust policy in a pre-specified policy class that can approximately maximize the smallest value of the set. Leveraging semi-parametric statistics, we develop an efficient policy learning method for estimating the defined robust optimal policy. A rate-optimal regret bound up to a logarithmic factor is established in terms of the number of trajectories and the number of decision points. Our regret guarantee subsumes the long-term average reward MDP setting as a special case and can be extended to the discounted infinite-horizon setting.

E0837: **Interpoint-ranking sign covariance to test independence**
Presenter: **Kehui Chen**, University of Pittsburgh, United States
Co-authors: Haeun Moon

An interpoint-ranking sign covariance is introduced, which is defined for general types of random objects with a meaningful similarity measure. We will show that it is zero if and only if the two random objects under consideration are independent. We will then introduce a test of independence based on the new interpoint-ranking sign covariance, and show that the proposed test is consistent against general types of alternatives. We will also present numerical experiments and data analysis to demonstrate the great empirical performance of the proposed method.

EO564 Room R14 ADVANCES IN STATISTICAL METHODS AND APPLICATION WITH DIGITAL DATA Chair: Shaoyang Ning

E0642: **Big-data infectious disease estimation: From flu to covid-19**
Presenter: **Shihao Yang**, Georgia Institute of Technology, United States

For epidemics control and prevention, timely insights of potential hot spots are invaluable. An alternative to traditional epidemic surveillance, which often lags behind real-time by days or even weeks, big data from the Internet provide important information about the current epidemic trends. We will present a few big-data approaches for influenza prediction, and how the approaches are applied to covid-19 prediction in the current pandemic.

E0811: **Recurrent event analysis in the presence of real-time high frequency data via random subsampling**
Presenter: **Walter Dempsey**, University of Michigan, United States

Digital monitoring studies collect real-time high-frequency data via mobile sensors in the subjects' natural environment. This data can be used to model the impact of changes in physiology on recurrent event outcomes such as smoking, drug use, alcohol use, or self-identified moments of suicide ideation. Likelihood calculations for the recurrent event analysis, however, become computationally prohibitive in this setting. Motivated by this, a random subsampling framework is proposed for computationally efficient, approximate likelihood-based estimation. A subsampling-unbiased estimator for the derivative of the cumulative hazard enters into an approximation of log-likelihood. The estimator has two sources of variation: the first due to the recurrent event model and the second due to subsampling. The latter can be reduced by increasing the sampling rate; however, this leads to increased computational costs. The approximate score equations are equivalent to logistic regression score equations, allowing for standard, "off-the-shelf" software to be used in fitting these models. Simulations demonstrate the method and efficiency-computation trade-off. We end by illustrating our approach using data from a digital monitoring study of suicidal ideation.

E0868: **The causal mediation analysis in the e-commerce industry**
Presenter: **Xuan Yin**, Etsy Inc, United States

Causal mediation analysis is a formal statistical framework to reveal the underlying causal mechanism in randomized experiments. The analysis has been widely employed in various disciplines. However, it has not been applied to online A/B tests, the online randomized experiments in the daily practice of the internet industry. Perhaps it is because online A/B tests in the internet industry are primarily for evaluation: estimating and testing the average treatment effect. We will discuss two of our recent works on the development of causal mediation analysis for producing insights for search and recommendation systems in the e-commerce industry. (1) Based on some evidence, it is hypothesized that search and recommendation systems could compete for users' attention, which leads to degradation in the overall performance of the website. We utilize causal mediation analysis to verify the hypothesis and quantify the competition formally. (2) It is common in the internet industry to develop algorithms offline to power online products that contribute to business KPIs. Evaluation metrics of algorithms are usually different from business KPIs. It is not clear which evaluation metric, among all available ones, should be the north star to guide the development of algorithms in order to optimize business KPIs. We extend causal mediation analysis and develop a novel approach, which is easy to implement and to scale up, to pick the north star.

E1087: **BETS: The dangers of selection bias in early analyses of the coronavirus disease (COVID-19) pandemic**
Presenter: **Nianqiao Ju**, Harvard University, United States
Co-authors: Qingyuan Zhao, Sergio Bacallado, Rajen D Shah

The coronavirus disease 2019 (COVID-19) has quickly grown from a regional outbreak in Wuhan, China, to a global pandemic. Early estimates of the epidemic growth and incubation period of COVID-19 may have been biased due to sample selection. Using detailed case reports from 14

locations in and outside mainland China, we obtained 378 Wuhan-exported cases who left Wuhan before an abrupt travel quarantine. We developed a generative model we call BETS for four key epidemiological events—Beginning of exposure, End of exposure, time of Transmission, and time of Symptom onset (BETS)—and derived explicit formulas to correct for the sample selection. We gave a detailed illustration of why some early and highly influential analyses of the COVID-19 pandemic were severely biased. All our analyses, regardless of which subsample and model were being used, point to an epidemic doubling time of 2 to 2.5 days during the early outbreak in Wuhan. A Bayesian nonparametric analysis further suggests that about 5% of the symptomatic cases may not develop symptoms within 14 days of infection and that men may be much more likely than women to develop symptoms within 2 days of infection.

EO085 Room R15 RECENT DEVELOPMENTS ON ANALYSIS OF NETWORKS Chair: Stefan Stein

E0256: **Community detection for hypergraph networks via regularized tensor power iteration**
Presenter: **Dong Xia**, Hong Kong University of Science and Technology, Hong Kong

To date, social network analysis has been largely focused on pairwise interactions. The study of higher-order interactions, via a hypergraph network, brings in new insights. We study community detection in a hypergraph network. We propose a new method for community detection that operates directly on the hypergraph. At the heart of our method is a regularized higher-order orthogonal iteration (reg-HOOI) algorithm that computes an approximate low-rank decomposition of the network adjacency tensor. Compared with existing tensor decomposition methods such as HOSVD and vanilla HOOI, reg-HOOI yields better performance, especially when the hypergraph is sparse. Given the output of tensor decomposition, we then generalize the community detection method SCORE from graph networks to hypergraph networks. We call our new method Tensor-SCORE. In theory, we introduce a degree-corrected block model for hypergraphs (hDCBM), and show that Tensor-SCORE yields consistent community detection for a wide range of network sparsity and degree heterogeneity. We apply our method to several real hypergraph networks which yield encouraging results. It suggests that exploring higher-order interactions provides additional information not seen in graph representations.

E0249: **A sparse beta model with covariates for networks**
Presenter: **Stefan Stein**, University of Warwick, United Kingdom
Co-authors: Chenlei Leng

Data in the form of networks are increasingly encountered in modern science and humanity. We present a new generative model, suitable for sparse networks commonly observed in practice, to capture degree heterogeneity and homophily, two stylized features of a typical network. The former is achieved by differentially assigning parameters to individual nodes, while the latter is materialized by incorporating covariates. Similar models in the literature often include as many nodal parameters as the number of nodes, leading to over-parametrization. As a result, they can only model relatively dense networks. For estimation, we propose the use of the penalized likelihood method with an ℓ_1 penalty on the nodal parameters, leading to a convex optimization formulation which immediately connects our estimation procedure to the LASSO literature. We highlight the differences of our approach to the LASSO method for logistic regression, emphasizing its feasibility to conduct inference for sparse networks, and study the finite-sample error bounds of the resulting estimator, as well as deriving a central limit theorem for the parameter associated with the covariates.

E0245: **Salient structure identification in complex networks by spectral periphery filtering**
Presenter: **Tianxi Li**, University of Virginia, United States
Co-authors: Ruizhong Miao

Complex networks have been intensively studied in the past fifteen years. In practice, the salient network structure of interest, instead of being directly observed, is often hidden in a larger network in which most structures are not informative. The noise and bias introduced by this overwhelming yet non-informative data can obscure the salient structure and limit the effectiveness of many network analysis methods. Traditionally, researchers treat this scenario as a core-periphery structure, and algorithms are designed to extract the core. Unfortunately, most of these methods rely on restrictive assumptions on both the core and the periphery components that seriously undermine their usefulness. We propose a random network model for the non-informative structure of networks without imposing a specific form for the core. Specifically, we assume that the non-informative nodes are connected to other nodes in a purely random pattern, while the core structure can take any informative pattern. Moreover, we propose an algorithm of core extraction. The algorithm is computationally efficient and comes with a theoretical guarantee of accuracy. We evaluated the proposed model in extensive simulation studies and also use it to extract core structures in a few real-world networks for downstream analysis.

E0866: **Network response regression for modeling population of networks with covariates**
Presenter: **Emma Jingfei Zhang**, University of Miami, United States

Multiple-network data are fast emerging in recent years, where a separate network over a common set of nodes is measured for each individual subject, along with rich subject covariates information. Existing network analysis methods have primarily focused on modelling a single network, and are not directly applicable to multiple networks with subject covariates. We present a new network response regression model, where the observed networks are treated as matrix-valued responses, and the individual covariates as predictors. The new model characterizes the population-level connectivity pattern through a low-rank intercept matrix and the parsimonious effects of subject covariates on the network through a sparse slope tensor. We formulate the parameter estimation as a non-convex optimization problem, and develop an efficient alternating gradient descent algorithm. We establish the non-asymptotic error bound for the actual estimator from our optimization algorithm. Built upon this error bound, we derive the strong consistency for network community recovery, as well as the edge selection consistency. We demonstrate the efficacy of our method through intensive simulations and two brain connectivity studies.

EO441 Room R16 RECENT ADVANCES IN ROBUST AND NONPARAMETRIC REGRESSION Chair: Yang Feng

E0254: **New regression model: Modal regression**
Presenter: **Weixin Yao**, UC Riverside, United States

Built on the ideas of mean and quantile, mean regression and quantile regression are extensively investigated and popularly used to model the relationship between a dependent variable Y and covariates x. However, the research about the regression model built on the mode is rather limited. We propose a new regression tool, named modal regression, that aims to find the most probable conditional value (mode) of a dependent variable Y given covariates x rather than the mean that is used by the traditional mean regression. The modal regression can reveal new interesting data structure that is possibly missed by the conditional mean or quantiles. In addition, modal regression is resistant to outliers and heavy-tailed data, and can provide shorter prediction intervals when the data are skewed. Furthermore, unlike traditional mean regression, the modal regression can be directly applied to the truncated data. Modal regression could be a potentially very useful regression tool that can complement the traditional mean and quantile regressions.

E0274: **Forward screening for high dimensional additive quantile regression**
Presenter: **Daoji Li**, California State University Fullerton, United States

A new feature screening approach is presented for high dimensional additive quantile regression. Under certain regularity conditions, we show that the proposed method all the important variables can be identified in a small number of steps. To remove noise variables after the screening step, we further implement variable selection via a modified Bayesian information criterion. We show that the smaller selected set still contains

all the important variables with overwhelming probability. The method and theoretical results are supported by several simulations and real data examples.

E0280: **Stock return predictability and cyclical movements in valuation ratios**
Presenter: **Li Chen**, Xiamen University, China
Co-authors: Deshui Yu, Difang Huang

According to present-value models, valuation ratios should predict future stock returns or cash flows but empirically show little power. The aim is to develop insights about stock return predictability, and to reconcile the contradicting findings. We decompose a valuation ratio into (i) a slow-moving component which reflects the local time-varying expected values of the valuation ratio as the results of persistent shocks, and (ii) a cyclical component which reflects rapid mean-reversion toward to the time-varying mean. The cyclical components of valuation ratios show statistically significant power for predicting future stock returns, both in-sample and out-of-sample, and the predictability is also economically significant. Conversely, the slow-moving components fail to predict returns, and therefore the components disguise the predictive information contained in valuation ratios for future returns. To ascertain our findings, we provide a direct line linking the decomposition approach to the present-value framework and the system of predictive regression.

E0375: **Learning non-smooth models: Instrumental variable quantile regressions and related problems**
Presenter: **Yinchu Zhu**, Brandeis University, United States

Computationally efficient methods are proposed that can be used for instrumental variable quantile regressions (IVQR) and related methods with statistical guarantees. We prove that the GMM formulation of IVQR is NP-hard, and finding an approximate solution is also NP-hard. We aim to obtain an estimate that has good statistical properties and is not necessarily the global solution of any optimization problem. The proposal consists of employing k-step correction on an initial estimate. The initial estimate exploits the latest advances in mixed-integer linear programming and can be computed within seconds. The theoretical contribution is that such initial estimators and Jacobian of the moment condition used in the k-step correction need not be even consistent and merely $k = 4\log n$ fast iterations are needed to obtain an efficient estimator. The overall proposal scales well to handle extremely large sample sizes because lack of consistency requirement allows one to use a very small subsample to obtain the initial estimate and the k-step iterations on the full sample can be implemented efficiently. We evaluate the performance of the proposal in simulations and an empirical example on the heterogeneous treatment effect of the Job Training Partnership Act.

EO201 Room R17 STATISTICS FOR HIGH-DIMENSIONAL HIGH-FREQUENCY DATA Chair: Markus Bibinger

E0272: **A shrinkage estimator of quadratic variation in high-dimensional settings**
Presenter: **Mikkel Slot Nielsen**, Columbia University, United States
Co-authors: Kim Christensen, Mark Podolskij

An estimator of the quadratic variation of high-dimensional semimartingales is proposed based on nuclear-norm penalization. Specifically, under suitable conditions, we prove a concentration inequality for estimators obtained by soft-thresholding of the eigenvalues of the realized variance. By relying on this result, we show that, by proper tuning, one can obtain an estimator of the quadratic variation which is minimax optimal up to a logarithmic factor and has the true rank with high probability. The theory is extended to include estimation of the local volatility and it is complemented by a simulation study as well as an empirical application.

E0756: **Rank tests for time-varying covariance matrices**
Presenter: **Lars Winkelmann**, Freie Universitaet Berlin, Germany

The model of a d-dimensional continuous-time martingale is considered. The process is observed under observational noise as is standard for microstructure noise models in high-frequency finance. We ask for testing the rank of the time-varying covariance matrix. The test problem is considered locally around some fixed point in time as well as uniformly and in mean over [0, 1]. The signal detection boundary, or optimal separation rate for which the test keeps power under H_1, is determined in all three cases. An interesting finding is that this rate does not only depend on the smoothness of the covariance matrix but also significantly on the spectral gap between adjacent eigenvalues. An application to the term structure of interest rates shows the practicability of the new tests.

E0750: **On Dantzig and Lasso estimators of the drift in a high dimensional Ornstein-Uhlenbeck model**
Presenter: **Dmytro Marushkevych**, University of Luxembourg, Luxembourg
Co-authors: Mark Podolskij, Gabriela Ciolek

New theoretical results are presented for the Dantzig and Lasso estimators of the drift in a high dimensional Ornstein-Uhlenbeck model under sparsity constraints. The focus is on oracle inequalities for both estimators and error bounds for several norms. We show that Dantzig and Lasso estimators have optimal rates, proving the restricted eigenvalue property solely under ergodicity assumption on the model. We also demonstrate the results of numerical analysis to uncover the finite sample performance of the Dantzig and Lasso estimators.

E0544: **A Bernstein-von Mises theorem for stochastic PDEs**
Presenter: **Randolf Altmeyer**, Cambridge University, United Kingdom

The Bayesian paradigm is considered in the context of statistics for SPDEs (stochastic partial differential equations). The goal is to perform nonparametric estimation of the diffusivity function in the stochastic heat equation, driven by space time white noise. Observations are given by local measurements, that is, the solution convoluted in a space with a compactly supported kernel function. Starting with a Gaussian process prior for the diffusivity, we discuss a Bernstein-von Mises theorem. This proves that the posterior distribution allows for asymptotically valid and optimal frequentist statistical inference on the diffusivity. As an important ingredient in the proof, we will also discuss the local asymptotic normality (LAN) property of the statistical model.

EO347 Room R18 RECENT DEVELOPMENTS OF COMPETING RISK DATA Chair: Feng-Chang Lin

E0509: **A nonparametric survival estimation method for dependent competing risk: An application in relative survival analysis**
Presenter: **Reuben Adatorwovor**, University of Kentucky, United States
Co-authors: Jason Fine

Quantifying disease-specific survival in patients with competing risk is generally done by disease-specific survival analysis when the cause of the event is known. Latent variable model is another method formulated for the unobserved event time. This approach may be the better one for population-based cancer survival studies because disease-specific survival estimates are invalid for the unreliable, misclassified or missing cause of death information. The cause of death due to disease competes with other causes of death, which creates a dependence between the event times. To relax the independence assumption, we formulate the dependence between the time to disease-specific death and the time to other causes of competing mortality using copula. A nonparametric copula-based methodology is used to fit the distributions of disease-specific death and other cause mortality using a function of the Kaplan-Meier estimator. Since the dependence structure for disease-related and other-cause mortality is unknown, we treat the copula as known with a sensitivity analysis conducted across a range of assumed dependence structures. We demonstrate the practical utility of our method through simulation studies with an application to French breast cancer data where we estimated the net and crude survival probabilities which are used for determining prognosis and treatment regimen for disease-specific survival.

E0512: **Classification of unknown cause of failure in competing risks: An application to recurrence of P.vivax malaria infection**
Presenter: **Yutong Liu**, University of North Carolina - Chapel Hill, United States
Co-authors: Feng-Chang Lin, Jessica Lin, Quefeng Li

A standard competing risks set-up requires both time-to-event and cause of failure to be fully observable for all subjects. However, in applications, the cause of failure may not always be observable, impeding the risk assessment. In some extreme cases, none of the causes of failure is observable. In the case of a recurrent episode of Plasmodium vivax malaria following treatment, the patient may have suffered a relapse from a previous infection or acquired a new infection from a mosquito bite. In this case, the time to relapse cannot be modeled when a competing risk, a new infection, is present. The efficacy of a treatment for preventing relapse from a previous infection may be underestimated when the true cause of infection cannot be classified. We developed a novel method for classifying the latent cause of failure under a competing risks set-up, which uses not only time to event information but also transition likelihoods between covariates at the baseline and at the time of event occurrence. Our classifier shows superior performance under various scenarios of simulation experiments. The method was applied to Plasmodium vivax infection data to classify recurrent malaria infections.

E1054: **Accounting for preinvasive conditions in analysis of invasive cancer risk: Application to breast cancer**
Presenter: **Jung In Kim**, The Pennsylvania State University, United States
Co-authors: Jason Fine, Shanshan Zhao

Ductal carcinoma in situ (DCIS), non-invasive cancer where abnormal cells have been found in the lining of the breast milk duct, is considered as the earliest stage of breast cancer. In epidemiology studies, there are several ways to deal with DCIS. DCIS cases are usually considered as censored cases by restricting the outcome only to invasive breast cancer. Alternatively, the first of either DCIS or invasive breast cancer is regarded as being the outcome. The former makes the strong assumption that DCIS and breast cancer are independent, while the latter fails to distinguish the risk of DCIS from that of breast cancer. In the Sister Study data, almost all women who had been diagnosed with DCIS were treated with lumpectomy or mastectomy, and they will not have invasive breast cancer after DCIS, clearly violating the independent censoring assumption. We propose a competing risks framework for analyzing breast cancer risks in the presence of DCIS by addressing the limitations of the conventional approaches. We demonstrate our approach via comprehensive simulation studies.

E1074: **Classification of disease progression via recurrent biomarkers using EM algorithm**
Presenter: **Huijun Jiang**, University of North Carolina at Chapel Hill, United States
Co-authors: Quefeng Li, Jessica Lin, Feng-Chang Lin

Many infectious diseases have more than one potential cause. The classification of infections from more than one possible cause is critical in effective disease control. Multistate model based on Markov processes is a typical approach to estimating the transition rate between the status of the disease. However, it can perform poorly when the problem of interest is the classification of unknown disease status. We aim to demonstrate that the transition likelihoods of disease biomarkers can be utilized to distinguish relapse from reinfection for malaria infection with high accuracy. A more general model for disease progression can be constructed to allow for additional disease states. We start from a multinomial logit model to estimate the disease transition probabilities and then utilize the transition information of disease biomarkers to provide a more accurate classification result. We apply the Expectation-Maximization (EM) algorithm for the estimation of unknown parameters, including the marginal probabilities of disease status. A comparison to the existing two-stage method shows that our classifier is consistent and has better accuracy, especially when the sample size is small. An application to data from 78 Cambodian P. vivax malaria patients is presented to demonstrate the practical use of our proposed method.

EO079 Room R19 STATISTICS IN NEUROSCIENCE I Chair: Jeff Goldsmith

E0787: **Mixed modeling frameworks for analyzing whole-brain network data**
Presenter: **Sean Simpson**, Wake Forest School of Medicine, United States

Brain network analyses have exploded in recent years, and hold great potential in helping us understand normal and abnormal brain function. Network science approaches have facilitated these analyses and our understanding of how the brain is structurally and functionally organized. However, the development of statistical methods that allow relating this organization to health outcomes has lagged behind. We have attempted to address this need by developing mixed-modeling frameworks that allow relating system-level properties of brain networks to outcomes of interest. These frameworks serve as a synergistic fusion of multivariate statistical approaches with network science, providing a needed analytic foundation for whole-brain network data. Here we delineate these approaches that have been developed for single-task, multitask, and dynamic brain network data.

E0844: **Harmonizing neuroimaging data acquired under complex study designs**
Presenter: **Russell Shinohara**, University of Pennsylvania, United States

As multi-center studies in imaging science become increasingly commonplace, there is a need for understanding and mitigating biases associated with the acquisition on multiple scanners. We will review the state-of-the-art in image harmonization, and explore harmonization in new settings such as longitudinal study designs and complex covariance structures in imaging features. We will conclude with a discussion of future directions and areas for improvement in study design and analysis.

E0920: **Online control of reach accuracy and why we need better functional data models for dynamic movement**
Presenter: **Julia Wrobel**, University of Colorado School of Public Health, United States

Reaching movements, as a basic yet complex motor behavior, are a foundational model system in neuroscience. In particular, there has been a significant recent expansion of investigation into the neural circuit mechanisms of reach behavior in mice. Nevertheless, quantification of mouse reach kinematics remains lacking. We quantitatively demonstrate the homology of mouse reach kinematics to primate reach, and also discover novel late-phase correlation structure that implies online control. Overall, the results highlight the declarative phase of reach as important in driving successful outcomes. Specifically, we develop and implement a novel statistical machine learning algorithm to identify kinematic features associated with successful reaches and find that late-phase kinematics are most predictive of outcome, signifying online reach control as opposed to pre-planning. Moreover, we identify and characterize late-phase kinematic adjustments that are yoked to mid-flight position and velocity of the limb, allowing for dynamic correction of initial variability, with head-fixed reaches being less dependent on position in comparison to freely-behaving reaches. Furthermore, consecutive reaches exhibit positional error-correction but not hot-handedness, implying opponent regulation of motor variability.

E1119: **The impact of autocorrelation in fMRI task and rest analysis**
Presenter: **Soroosh Afyouni**, University of Oxford, United Kingdom
Co-authors: Thomas Nichols

Time series obtained using fMRI are notoriously autocorrelated. Although the source of the autocorrelation is not completely known, the dependency between the observations induced by autocorrelation violates the assumption behind the majority of conventional statistical methods used in rest and task fMRI analysis. We show that in resting-state functional connectivity, the variance of Pearson correlations, the most widely used measure of connectivity on subject level, is excessively biased due to autocorrelation. We propose a novel variance estimator for sample correlation coefficients which accounts for such dependencies. Further, we show that the existing methods for accounting autocorrelation in noise for subject-

level task fMRI come short in modern rapidly sampled fMRI scans. Using a combination of spectral methods, we propose a whitening technique which successfully flattens the spectrum of the noise across various task paradigms in fMRI acquisitions above 1Hz.

EO083 Room R20 CAUSAL INFERENCE AND GRAPHICAL MODELS Chair: Xavier de Luna

E0829: **Contrasting identification criteria of average causal effects: Asymptotic variances and semiparametric estimators**
Presenter: **Tetiana Gorbach**, Umea University, Sweden
Co-authors: Xavier de Luna, Juha Karvanen, Ingeborg Waernbaum

The back-door criterion (based on pre-treatment covariates) and the front-door criterion (based on mediators) are commonly used to identify an average causal effect from observational data given that the data generating mechanism is compatible with a directed acyclic graph (DAG). Even if the back- and the front-door criteria are not fulfilled, the causal effect might also be identified using mediators and pre-treatment covariates together when a condition that we call the two-door criterion holds. When several criteria hold for the DAG at hand, one may want to choose the criterion that provides the most efficient estimator. We give theoretical and numerical comparisons of asymptotic variances of semiparametric estimators based on the back-, the front-, and the two-door identification assumptions when any of the criteria hold simultaneously. The theoretical and simulation results obtained show that none of the criteria systematically yields the lowest asymptotic variance, or in other words, no estimation strategy is going to be most efficient in all situations. We, however, give conditions under which the two-door criterion is known to outperform the back-door criterion.

E0650: **A potential outcomes calculus for identifying conditional path-specific effects**
Presenter: **Daniel Malinsky**, Columbia University, United States
Co-authors: Ilya Shpitser, Thomas Richardson

The do-calculus is a well-known deductive system for deriving connections between interventional and observed distributions. It has been proven complete for several important identifiability problems in causal inference. Nevertheless, as it is currently defined, the do-calculus is inapplicable to causal problems that involve complex nested counterfactuals which cannot be expressed in terms of the do operator. Such problems include analyses of path-specific effects and dynamic treatment regimes. We present the potential outcome calculus (po-calculus), a natural generalization of do-calculus for arbitrary potential outcomes. We thereby provide a bridge between identification approaches which have their origins in artificial intelligence and statistics, respectively. We use po-calculus to give a complete identification algorithm for conditional path-specific effects with applications to problems in mediation analysis and algorithmic fairness.

E0830: **Efficient conditional instrumental set selection**
Presenter: **Leonard Henckel**, ETH Zurich, Switzerland
Co-authors: Marloes Maathuis

Instrumental variable estimators are a popular tool for causal effect estimation in the presence of unmeasured or latent confounding. However, it is well known that they tend to suffer from low accuracy. We consider ways to improve the 2SLS estimator's accuracy by improving the conditional instrumental set (CIS) selection. Presupposing knowledge of the underlying causal structure in the form of an acyclic directed mixed graph (ADMG), we develop three graphical tools to aid in the selection of more efficient CISs. First, we reformulate the asymptotic variance formula for the 2SLS estimator in a way that in particular provides new insights into how the choice of conditioning set, commonly referred to as the exogenous variables, affects the asymptotic variance. Second, we derive a graphical criterion allowing us to compare the asymptotic variance of certain pairs of CISs. Third, we construct a near-optimal valid CIS, that is, a CIS with an efficiency guarantee that cannot be improved without additional non-graphical information. We also use the first two results to derive guidelines that are helpful even in the absence of precise graphical knowledge and can be applied using only instrumental validity checks.

E0484: **On the monotonicity of a binary confounder**
Presenter: **Jose M Pena**, Linkoping University, Sweden

The focus is on the average causal effect of a binary treatment on an outcome when a binary confounder confounds this relationship. Suppose that the confounder is unobserved, but a nondifferential proxy of it is observed. We will show that under certain monotonicity assumption that is empirically verifiable, adjusting for the proxy produces a measure of the effect that is between the unadjusted and the true measures. We will also show through experiments that most random parameterizations result in a proxy-adjusted effect that lies between the unadjusted and the true ones. However, only half of them satisfy the monotonicity condition named above. Therefore, the condition is sufficient but not necessary. This result should be interpreted with caution because we are seldom interested in a random parameterization. Therefore, we will also discuss some nonmonotonic cases (albeit empirically untestable) where the proxy-adjusted effect still lies between the unadjusted and the true ones.

EO654 Room R21 RECENT STATISTICAL ADVANCES IN HIGH-DIMENSIONAL BIOMEDICAL APPLICATIONS Chair: Subharup Guha

E0947: **Connectivity regression**
Presenter: **Jeffrey Morris**, University of Pennsylvania, United States
Co-authors: Veerabhadran Baladandayuthapani, Neel Desai

An important problem posed by modern big data is the regression of multivariate associations on predictors. One example in neuroimaging involves functional connectivity, discerning associations among brain regions and assessing how these vary according to discrete and continuous factors. We introduce a general connectivity regression framework that can determine which factors impact connectivity and characterize which graph edges vary by each significant factor. Our approach involves projecting the subject-specific connectivity estimates into an alternative space for which Gaussian assumptions are justified and positive definiteness in the original space in ensured, and in which we perform multivariate regression using a multivariate-spike-and-slab lasso to simultaneously perform variable selection on covariate effects on edges and detect edge-to-edge associations. This penalty increases efficiency in estimation and covariate selection through the principles of seemingly unrelated regressions, and we demonstrate small sample properties by simulation. We apply this method to data from the Human Connectome Project and discuss the generality and extendibility of the framework.

E0968: **Efficient estimation of SNP heritability using Gaussian predictive process in large scale cohort studies**
Presenter: **Saonli Basu**, University of Minnesota, United States
Co-authors: Souvik Seal, Abhirup Datta

For decades, Linear Mixed Model (LMM) has been the most popular tool for estimating heritability in twin and family studies. Recently, with the advent of high throughput genetic data, there is quite a bit of interest to estimate heritability by using a high-dimensional Genetic Relationship Matrix (GRM) constructed from genome-wide SNP data on distantly related individuals. Fitting such an LMM in large scale cohort studies, however, is tremendously challenging due to high dimensional linear algebraic operations. We simplify the LMM unifying the concepts of Genetic Coalescence and Gaussian Predictive Process modeling, greatly alleviating the computational burden. The method, named as PredLMM, has much better computational complexity than most of the existing packages and thus, provides an efficient alternative of estimating heritability in large scale cohort study. We illustrate our approach with extensive simulation studies and use PredLMM to estimate the heritability of multiple quantitative traits from the UK Biobank cohort.

E0755: **Summarizing posterior clustering distributions**
Presenter: **David Dahl**, Brigham Young University, United States
Co-authors: Devin Johnson, Peter Mueller
The aim is to address the problem of point estimation of a clustering distribution based on posterior samples, as well as the assessment of clustering uncertainty. We both extend the literature of loss functions for Bayesian clustering and also introduce a fast, scalable optimization procedure to obtain an optimal Bayesian estimate. Our approach is a stochastic search based on a series of micro-optimizations performed in random order and is embarrassingly parallel. We explain the algorithm and computational shortcuts, demonstrate the software, and compare its performance against increasingly-challenging applications of Bayesian clustering.

E0967: **Identifying cancer driver genes from differential co-expression networks**
Presenter: **Tyler Grimes**, University of North Florida, United States
The underlying driver of cancer stems from somatic mutations in the genome that change the function of gene products. However, not all mutations are associated with cancer progression. Rather, such "passenger" mutations are a symptom of DNA instability. Only a small portion of mutations are actual "drivers" that are responsible for disease progression. A methodology is proposed for analyzing gene expression data, to identify driver genes by considering both the functional changes of genes and the clinical relevance of those changes. Functional changes are identified by performing a differential network analysis, which compares the structure of gene-gene associations across different stages of cancer. Genes that are differentially connected indicate a change in their functional activity along with cancer progression. Clinical relevance of these differential connections is assessed using a survival model to predict overall survival. Potential driver genes are identified for Neuroblastoma and Breast cancer patient populations. We identify regulatory pathways that are both differentially connected and whose expression profile is predictive of overall survival. We plan to analyze additional cancers from the TCGA data repository and perform a meta-analysis to identify any pan-cancer driver genes or biomarkers.

EO091 Room R22 BAYESIAN MODEL COMPARISON Chair: Mattias Villani

E0796: **The good, the bad, and the ugly: Overconfident Bayesian model selection in molecular phylogenetics**
Presenter: **Ziheng Yang**, University College London, United Kingdom
Co-authors: Tianqi Zhu
Bayesian model selection is widely used to compare species phylogenetic trees in molecular phylogenetics. It is noted to produce high and spurious posterior probabilities for phylogenies in large datasets, but the precise reasons for this overconfidence are unknown. The empirical observations have prompted us to study the asymptotic behavior of Bayesian model selection when the models under comparison have the same number of parameters and are equally wrong. We found that in such cases, Bayesian model selection exhibits surprising and polarized behaviors in large datasets, supporting one model with full force in some datasets while supporting another in others. If one model is slightly less wrong than the other, the less wrong model will eventually win when the amount of data increases, but the method tends to become overconfident before it becomes reliable. This extreme behavior appears to be part of the reason for the spuriously high posterior probabilities for evolutionary trees. We discuss a few strategies suggested in the literature, but the question of "what should one do" remains open.

E0672: **When Bayesian model probabilities are overconfident**
Presenter: **Oscar Oelrich**, Deparment of Statistics, Stockholm University, Sweden
Co-authors: Mattias Villani, Mans Magnusson, Shutong Ding, Aki Vehtari
Bayesian model comparison is often based on the posterior distribution over the set of compared models. This distribution is often observed to concentrate on a single model even when other measures of model fit or forecasting ability indicate no strong preference. Furthermore, a moderate change in the data sample can easily shift the posterior model probabilities to concentrate on another model. We document overconfidence in two high-profile applications in economics and neuroscience. To shed more light on the sources of overconfidence, we derive the sampling variance of the Bayes factor in univariate and multivariate linear regression. The results show that overconfidence is likely to happen when i) the compared models give very different approximations of the data-generating process, ii) the models are very flexible with large degrees of freedom that are not shared between the models, and iii) the models underestimate the true variability in the data.

E0670: **Using stacking to combine Bayesian predictive distributions**
Presenter: **Yuling Yao**, Columbia University, United States
A general challenge in statistics is the prediction in the presence of multiple candidate models or learning algorithms. Bayesian model averaging is flawed in the M-open setting in which the true data generating process is not one of the candidate models being fit. Equipped with a proper scoring rule, stacking is a better approach to combine Bayesian predictive distributions. It yields asymptotically optimal predictions for future data and is more desired than single model selection. Furthermore, stacking can be used to combine posterior draws of one model and give better predictive performance than full Bayesian inference under misspecified models. Finally, we present that stacking can be combined with hierarchical modeling in structured data.

E1067: **Uncertainty in Bayesian leave-one-out cross-validation based model comparison**
Presenter: **Aki Vehtari**, Aalto University, Finland
Co-authors: Tuomas Sivula, Mans Magnusson
Leave-one-out cross-validation (LOO-CV) is a popular method for comparing Bayesian models based on their estimated predictive performance on new, unseen, data. Estimating the uncertainty of the resulting LOO-CV estimate is a complex task, and it is known that the commonly used standard error estimate is often too small. We analyse the frequency properties of the LOO-CV estimator and study the uncertainty related to it. We provide new results of the properties of the uncertainty both theoretically and empirically and discuss the challenges of estimating it. We show that problematic cases include: comparing models with similar predictions, misspecified models, and small data. In these cases, there is a weak connection in the skewness of the sampling distribution and the distribution of the error of the LOO-CV estimator. We show that it is possible that the problematic skewness of the error distribution, which occurs when the models make similar predictions, does not fade away when the data size grows to infinity in certain situations.

EO470 Room R23 BAYESIAN CAUSAL MODELLING OF TREATMENT STRATEGIES Chair: Erica Moodie

E1016: **Adaptive treatment allocation and selection in multi-arm clinical trials: A Bayesian perspective**
Presenter: **Elja Arjas**, University of Oslo, Norway
Clinical trials are an instrument for making informed decisions. Here we consider adaptive designs mainly from the perspective of multi-arm Phase II clinical trials, in which one or more experimental treatments are compared to a control. The same ideas can be applied, essentially without change, in confirmatory Phase III trials, where only a single experimental treatment is compared to a control. Still, the planned size of the trial is larger. In both situations, treatment allocation of individual patients is assumed to take place according to a fixed block randomization, albeit with an important twist: The performance of each treatment arm is assessed after every measured outcome, in terms of the posterior distribution of a corresponding model parameter. Different treatments arms are then compared to each other according to pre-defined criteria. If a treatment arm in

such a comparison is found to be sufficiently clearly inferior to the currently best candidate, it can be closed off either temporarily or permanently from further patient accrual.

E0772: **Agent-based modeling for medical research: Economic impact of generic antiretrovirals in France for HIV patients care**
Presenter: **Nicolas Savy**, Toulouse Institute of Mathematics, France
Co-authors: Philippe Saint-Pierre

Agent-based modeling consists of a set of models whose aim is to mimic the behavior of individuals in a random environment. In a health context, agent-based modeling may be used to simulate the behavior of patients or the effect of a treatment on patients. To do so, virtual patients are randomly generated, and models are used to predict their medical outcomes under different scenarios of treatment or treatment effects. By comparing these scenarios, it is possible to derive an estimate of the effect of an intervention or treatment on medical outcomes. That strategy is usually called In Silico Clinical Trial (ISCT). We will highlight the milestones of this strategy, which is strongly based on predictive models. Artificial Intelligence has shown wonderful power for prediction, although many strategies are available to predict an outcome from data. We focus our attention on the main methodological pitfall: to perform agent-based modeling, a predictive model is not enough, sharp modeling of the error of prediction is necessary. As an example, we will discuss an agent-based model developed to simulate patient trajectories and treatment use over a five years period. Comparing the cost results obtained for trajectories simulated under different predefined scenarios then allows us to build a Budget Impact Model as well as sensitivity analyses on several parameters of importance.

E0950: **Bayesian semiparametric inference and selection for dynamic treatment regimes**
Presenter: **David Stephens**, McGill University, Canada

Computational strategies are developed that allow fully Bayesian inference to be carried out for possibly misspecified models. Behaviour under, and robustness to, misspecification is widely studied in the frequentist world but is not prominent amongst Bayesians. We will demonstrate how the computational approaches provide exact Bayesian inference and how this can be deployed in the setting of dynamic treatment regimes.

E0771: **Subgroup analysis with time to event outcomes**
Presenter: **Peter Mueller**, UT Austin, United States
Co-authors: Satoshi Morita, Hiroyasu Abe

A utility-based Bayesian approach to population finding and subgroup analysis is discussed. The approach casts the population finding process as a formal decision problem together with a flexible probability model using a flexible model, such as random forests or other non-parametric Bayesian models, to fit the data. In contrast, the decision is constrained to be parsimonious and interpretable. We define a utility function that addresses the competing aims of the desired report. We illustrate the approach with a joint time-to-event and toxicity outcome for subgroup analysis, and with a time-to-event outcome in the context of an umbrella trial master protocol.

EO349 Room R24 NEW DEVELOPMENTS IN SURVEY SAMPLING Chair: David Haziza

E0263: **Model-assisted estimation through random forests in finite population sampling**
Presenter: **Mehdi Dagdoug**, Universite de Bourgogne Franche-Comte, France
Co-authors: Camelia Goga, David Haziza

Estimation of finite population parameters is of primary interest in survey sampling. At the estimation stage, auxiliary information is often available for all population units. The model-assisted approach uses this supplementary source of information to construct estimators. We propose new classes of model-assisted estimators based on random forests. Generally speaking, random forest is an ensemble method which consists of creating a large number of regression trees and combining them to produce more accurate predictions than a single regression tree would. Under mild conditions, the proposed model-assisted estimators are shown to be asymptotically design unbiased and consistent. A consistent variance estimator is proposed. The asymptotic distribution of the estimators is obtained, allowing for the use of confidence intervals. Simulations illustrate that the proposed estimator is efficient and can outperform state-of-the-art estimators, especially in complex settings.

E0882: **Two-step indirect sampling with application to the French postal traffic**
Presenter: **Estelle Medous**, University of Toulouse 1, France
Co-authors: Camelia Goga, Anne Ruiz-Gazen, Jean-Francois Beaumont, Alain Dessertaine, Pauline Puech

In surveys, there may be no sampling frame for the target population. A solution is to find a frame population linked in some way to the target population and use indirect sampling to draw a sample in the target population. The sampling weights of the sampled units can be determined using the General Weight Share Method (GWSM). This method cannot be applied when some of the links between the frame population and the sample in the target population are missing. A solution to avoid this issue is to consider an intermediate population linked in some way to both the frame and target populations. Then the GWSM can be used twice, first between the frame and intermediate populations and then between the intermediate and target populations. As illustrated with the French postal traffic survey, this double GWSM appears to be deteriorating the precision of estimators in some situations. Mathematical expressions for the loss of precision can be derived exactly for some sampling designs such as Poisson sampling or Simple Random Sampling Without Replacement. Using these mathematical expressions, it is possible to highlight the magnitude of the loss of precision in some practical situations.

E0960: **Donor imputation for multivariate missing data**
Presenter: **Audrey-Anne Vallee**, Universite Laval, Canada
Co-authors: Yves Tille, Esther Eustache

Swiss cheese nonresponse, also known as non-monotone nonresponse, occurs when every variable of a survey contains missing values without a particular pattern. The estimators of the parameters of interest can be considerably affected by the missing values which introduce a bias and an increase in the variability. To reduce the effects of nonresponse, the missing values are usually imputed. When several variables of a dataset need to be imputed, it may be difficult to preserve the distributions and the relations between the variables. Balanced K-nearest neighbor imputation is extended to treat Swiss cheese nonresponse. The method uses random imputations by donors, and it is constructed to meet the following requirements. First, a nonrespondent should be imputed by neighboring donors. Next, all missing values of a nonrespondent should be imputed by the same donor. Last, the donors are selected in order to satisfy some balancing constraints allowing to decrease the variance of the estimators. To meet all the requirements, a matrix of imputation probabilities is constructed using calibration techniques. The donors are then selected with these imputation probabilities and balanced sampling methods.

E0824: **Controlling the bias for M-quantile estimators for small area**
Presenter: **Francesco Schirripa Spagnolo**, Universita di Pisa, Italy
Co-authors: Gaia Bertarelli, Raymond Chambers, David Haziza, Nicola Salvati

When representative outlier units are a concern for estimation of population quantities, it is essential to pay attention to them in a small area estimation (SAE) context, where sample sizes are very small and the estimation in often model-based. Standard approaches use plug-in robust prediction replacing parameter estimates in optimal but outlier-sensitive predictors by outlier robust versions (robust-projective approach). These predictors are efficient under the correct model. Still, they may be sensitive to the presence of outliers because they use plug-in robust prediction, which usually leads to a low prediction variance and a considerable prediction bias. DA bias correction method for models with continuous response variables has been proposed. We apply two general methods to reduce the prediction bias of the robust M-quantile predictors in SAE. The first

estimator is based on the concept of conditional bias and extends previous results. The second one is a unified approach to M-quantile predictors based on a full bias correction. A Monte-Carlo simulation study is conducted. Results confirm that our approaches improve the efficiency and reduce the prediction bias of M-quantile predictors when the population contains units that may be influential if selected in the sample. Data from the EU-SILC 2017 survey in Italy are analysed.

EO387 Room R25 BRANCHING PROCESSES: THEORY, COMPUTATION AND APPLICATIONS I Chair: Miguel Gonzalez Velasco

E0818: **A predator-prey two-sex branching process**
Presenter: **Carmen Minuesa Abril**, Autonomous University of Madrid, Spain
Co-authors: Cristina Gutierrez Perez

The first branching process to describe the interaction of predator and prey populations with sexual reproduction is presented. We consider a two-type branching process, where the first type corresponds to the predator population and the second one to the prey population. The interaction and survival of both groups are modelled through control functions depending on the current number of individuals of each type in the ecosystem. The resulting model is a two-type two-sex controlled branching model. Due to the variety of mating systems, we focus on the promiscuous mating, where each female mates with only one male, whenever there are males in the population, but the same male could mate with more than one female. In this context, and given their interest for the conservation of species, we provide necessary and sufficient conditions for the ultimate extinction of both species, the fixation of one of them and the coexistence of both of them.

E0766: **Branching processes in varying environment**
Presenter: **Sandra Palau**, IIMAS, UNAM, Mexico

Branching processes and its extension are studied when the offspring distribution is changing over time. The extinction probability is analyzed. By using a two spine decomposition, the law of the process conditioned on survival will be given.

E0975: **Numerical schemes and algorithms for branching processes in modelling Coronavirus (COVID19) pandemics**
Presenter: **Maroussia Slavtchova-Bojkova**, Sofia University, Bulgaria

A new simulation methodology oriented to model the spread of the COVID19 pandemic caused by SARS-CoV-2 coronavirus is developed. There are many complications when modelling an outbreak of a novel infectious disease. To address some of these, we have described a possible technique to serve as part of a generally applicable toolkit. The mutual concern of estimation and simulation efforts is critical. Our methodology is based on the general branching models, which turned out to be more appropriate and flexible for describing the spread of an infection in a given population, than discrete-time ones. Concretely, Crump-Mode-Jagers branching processes are considered as proper candidates of infectious diseases modelling with incubation period like measles, mumps, avian flu, etc. and including the newly immerged COVID19 pandemic, as well. It can be pointed out that the developed methodology applies to the diseases that follow the so-called SIR (susceptible-infected-removed) and SEIR (susceptible exposed-infected-removed) scheme in terms of epidemiological models. Different forecasts are proposed and compared on the ground of real data and simulation examples.

E1018: **Bayesian estimation of controlled branching process choice without explicit likelihood**
Presenter: **Ines M del Puerto**, University of Extremadura, Spain
Co-authors: Miguel Gonzalez Velasco, Carmen Minuesa Abril

The purpose is to approximate the posterior distribution of the parameters of interest of controlled branching processes without explicit likelihood calculations nor any knowledge of the maximum number of offspring that an individual can produce. We consider that only the population sizes at every generation and at least the number of progenitors of the last generation are observed. Still, the number of offspring that every individual gives birth to is unknown at any generation. The method proposed is two-fold. We firstly make use of an Approximate Bayesian Computation based on sequential Monte Carlo (SMC ABC) model choice algorithm to estimate the posterior distribution of the maximum reproductive capacity. Secondly, to estimate the posterior distribution of the parameters of interest, we run the rejection ABC algorithm and the post-processing on the output of the previous method by considering an appropriated summary statistic. The accuracy of the proposed method is illustrated employing a simulated example developed with the statistical software R.

CO033 Room R02 TIME SERIES ECONOMETRICS MEETS CROSS SECTIONAL HETEROGENEITY Chair: Etsuro Shioji

C0810: **Fiscal multipliers in the most aged country: Empirical evidence and theoretical interpretation**
Presenter: **Hiroshi Morita**, Hosei University, Japan

The purpose is to investigate how population ageing impacts the effectiveness of a government spending shock. We estimate a panel VAR model with prefectural data in Japan, the world's fastest ageing country and reveal that a government spending shock becomes less effective as the ageing rate increases. Subsequently, we construct a New Keynesian model with workers and retirees, which can replicate our empirical findings. This highlights the role of the supply-side channel through which workers facing a liquidity constraint can benefit from increased disposable income, in generating the state-dependent effect of the government spending shock. The theoretical finding may suggest that promoting labour market participation by elderly people could increase the effectiveness of a government spending shock amid a rapidly ageing society.

C1043: **Asymmetric risk sharing and the business cycle**
Presenter: **Haerang Park**, Seoul National University, Korea, South
Co-authors: Soyoung Kim

Using panel data of 212 countries from 1970 to 2018, we provide evidence on asymmetric international risk sharing over the business cycle. Negative local shocks are amplified by lower international risk sharing, whereas negative global shocks are transmitted to nations around the world through greater international risk sharing. The patterns are not observed in positive shocks. The cyclicality is driven largely by income channel in local recessions due to rising borrowing costs, and by credit market channel in global recessions with a heightened precautionary saving motive. Consumption would be larger in the absence of the cyclical patterns in international risk sharing. More diversified cross-ownership of assets and international transfers could mitigate cyclical risk sharing and its negative impact on consumption. Cyclicality in international risk sharing has yet to be well examined in the international economics literature. Previous studies investigate it with limited cross-sectional variations and obtain mixed results. We provide more robust evidence based on a large cross-country dataset and identify channels of risk sharing that determine the cyclicality.

C0871: **Daily dynamics of retail gasoline price dispersion in Japan**
Presenter: **Etsuro Shioji**, Hitotsubashi University, Japan
Co-authors: Shiro Yuasa

The evolution of price distribution across gasoline stations in Japan is investigated. The data is obtained from a price comparison site and covers all the days between August 2018 and early January 2020. We control for various features of individual stations, such as regions, corporate group affiliations, and types of extra services offered at those stations. We also adjust for the infamously heavy Japanese taxes on gasoline. After all these treatments, the skewness and the kurtosis of the price distribution turn out to be relatively stable over time. On the other hand, we find that the price dispersion across the shops tends to widen when the world price of crude oil goes down, a tendency that is similar to those reported for other countries by previous studies. We study causes behind this pattern.

C0895: **Identifying and estimating the long-run effect of income distribution on the aggregate consumption**
Presenter: **Yoosoon Chang**, Indiana University, United States
Co-authors: Joon Park, Changsik Kim, Hwagyun Kim
The aim is to identify and estimate the long-run effect of income distribution on aggregate consumption. Permanent components of income and consumption are obtained by functional Beveridge-Nelson decomposition of U.S. Consumer Expenditure Survey data. From the permanent income distribution, we identify two factors, the level (aggregate) and the spread (redistribution), that affect permanent consumption. Longrun consumption is most positively affected by households with monthly earnings of around 2,000 dollars, households with lower income have negative effects on aggregate consumption, and those with 5,000 dollars or more respond little to income redistribution. Limited income sharing across households, high entry barriers, and nontrivial adjustment costs associated with both human and physical capital accumulation may contribute to the empirical findings. Taking the estimated long-run response function as the optimal behavior of households, counterfactual taxation exercises suggest that purely redistributive policies can increase the permanent component of aggregate consumption by 250%.

CO161 Room R03 SEMI- AND NONPARAMETRIC REGRESSION FOR TIME SERIES AND PANEL DATA Chair: Harry Haupt

C1134: **Measuring macroeconomic convergence and divergence within emu using long memory**
Presenter: **Theoplasti Kolaiti**, Leibniz University Hannover, Germany
Co-authors: Lena Draeger, Philipp Sibbertsen
Early studies before the start of EMU demonstrate some success in terms of nominal convergence of the member states. In contrast, others use cointegration analysis to demonstrate potential long-run stability problems with respect to macroeconomic dynamics in the so-called periphery countries. The aim is to measure the convergence or divergence of EMU inflation rates and industrial production by using several semiparametric methods to test for the existence of fractional cointegration relations. The notion of fractional cointegration allows for long-term equilibria with a higher degree of persistence than allowed for in the standard cointegration framework. We investigate both inflation and industrial production of EMU countries beginning with the introduction of the common currency and including the financial crisis and post-crisis period. Core, as well as periphery countries, are included in the study. By modelling possible breaks in the persistence structure, we find evidence of fractional cointegration as well as a lower persistence before the crisis and a higher persistence by less evidence for fractional cointegration during the crisis. A second break which indicates the end of the crisis can be found as well. In addition, higher inflation persistence can be found for periphery than for core countries.

C1117: **A semiparametric panel data model with common factors and spatial dependence**
Presenter: **Alexandra Soberon**, Universidad de Cantabria, Spain
Co-authors: Juan Manuel Rodriguez-Poo, Antonio Musolesi
New semiparametric heterogeneous panel data models are proposed which handle complex and relevant empirical problems, simultaneously: (i) functional misspecification by modelling stochastic observed common factors with a nonparametric function instead of assuming the usual parametric form; (ii) cross-sectional dependence originated simultaneously from common factors and spatial dependence, from the latter neither imposing a specific parametric spatial diffusion process nor requiring the specification of a given interaction matrix, but being directly derived from the data; iii) heterogeneous relations. We first propose a new estimation that extends the common correlated effect (CCE) approach to such a semiparametric spatially augmented framework. Then, Generalized Least Squares (GLS)-type estimators improving efficiency are proposed by taking into account the dependence structure. Asymptotic normal distributions are derived when the time dimension is large while the cross-sectional dimension need not be. Small sample properties of the estimators are investigated by Monte Carlo experiments, and an empirical application on the knowledge capital production function is conducted.

C1155: **Estimating change points in nonparametric time series regression models**
Presenter: **Leonie Selk**, Helmut-Schmidt-University, Germany
A regression model is considered that allows for time-series covariates as well as heteroscedasticity with a regression function that is modelled nonparametrically. We assume that there exists a change point such that the regression function changes at the unknown time $\lfloor ns_0 \rfloor$, $s_0 \in [0,1]$, and our aim is to estimate the (rescaled) change point s_0. The considered estimator is based on a Kolmogorov-Smirnov functional of the marked cumulative sum of residuals. We show the consistency of the estimator and prove a rate of convergence of $O_P(n^{-1})$. Additionally, we investigate the case of lagged dependent covariates, that is, autoregression models with a change in the nonparametric (auto-) regression function and give a consistency result. The method of proof also allows for different kinds of functionals such that Cramér-von Mises type estimators can be considered similarly. Finite sample simulations indicate the good performance of our estimator in regression as well as autoregression models and a real data example shows its applicability in practice.

C0170: **CLT for dependent heterogenous processes and applications to generalized quantile regression**
Presenter: **Harry Haupt**, University of Passau, Germany
A new CLT is proposed for weakly dependent heterogeneous processes, which is adequate for deriving the limiting distribution of L1-norm estimators. As an application of the main result, we simplify and extend previous results on the asymptotic normality of generalized and nonlinear quantile regression estimators in a semiparametric setting.

CO093 Room R06 CLIMATE CHANGE ECONOMETRICS AND FINANCIAL MARKETS Chair: Luca De Angelis

C0386: **A time-varying Greenium for European stocks**
Presenter: **Lucia Alessi**, European Commission - Joint Research Centre, Italy
Co-authors: Elisa Ossola, Roberto Panzica
The aim is to study the evolution of the Greenium, i.e. a risk premium linked to firms' greenness and environmental transparency, based on individual stock returns. The Greenium is associated with ith a priced 'greenness and transparency' factor, which considers both companies' greenhouse gas emissions and the quality of their environmental disclosures. By estimating an asset pricing model with time-varying risk premia, we analyze the evolution of the European Greenium from January 2006 to December 2019. We show that the Greenium dropped after the Paris Agreement was reached in December 2015, indicating investors' willingness to earn a lower return, ceteris paribus, to hold greener and more transparent stocks. The Greenium started to increase again at the end of 2016, with the election of Donald Trump.

C0431: **On the relationship between corporate environmental responsibility and financial performance: A portfolio analysis**
Presenter: **Thomas Alexopoulos**, University of Peloponnese , Greece
Co-authors: Elettra Agliardi
In recent years there has been a significant influx of environmentally responsible oriented investors. Climate change is the leading cause and the most important factor for asset managers considering sustainable investments. Our research uses the environmental pillar of ESG as a proxy for environmental corporate responsibility. We examine the performance of environmentally clustered portfolios by using simple quantitative investment strategies with optimum asset rotation. Post-hoc, sample-split analysis with non-parametric tests has been performed. The Paris Agreement is additionally examined as a shock event. The results suggest that environmental status is a key characteristic for divergent financial behaviors and that neither the environmental leaders nor the laggards earn the most. Supporting evidence of positive spillovers in high Environmental Clusters

are identified after the Paris Agreement. We also discuss the potential implications both from the perspective of potential investors and of future growth of a clean energy-related class of assets.

C0418: **Climate change awareness: Empirical evidence for the European Union**
Presenter: **Claudio Morana**, Universita di Milano Bicocca, Italy

Public attitudes on climate change in Europe over the last decade are assessed. Using aggregate figures from the Special Eurobarometer surveys on Climate Change, we find that the evolution of climate change attitudes over time is well described by the "S-shaped" information dissemination model, conditional to various socioeconomic and climatological factors. In particular, we find that environmental concern is directly related to per capita income, social trust, secondary education, the physical distress associated with hot weather, and loss caused by extreme weather episodes. It is also inversely related to greenhouse gas emissions and tertiary education. Moreover, we find a significant, opposite impact for two temporal dummies for years 2017 and 2019, which, consistent with their timing and expected role for opinion leaders, is tempting to associate with Donald Trump's denial campaigns and the U.S. Paris Agreement withdrawal announcement and Greta Thunberg's environmental activism, respectively.

C0417: **The impact of compounding COVID-19 and climate risk on sovereign debt**
Presenter: **Luca De Angelis**, University of Bologna, Italy
Co-authors: Irene Monasterolo, Anja Duranovic

A country-specific Structural VAR (SVAR) is used to measure the impact of COVID-19 pandemic and climate-related disasters on the sovereign bond risk of the Caribbean countries. In particular, we model the CDS spread change, the bond spread change, the number of deaths by COVID, and one or two variables for climate-related disasters as a vector of endogenous variables. Then, a structural (Cholesky) analysis is put forward to investigate the impact of a shock in climate or pandemic or both (compound risk).

CO474 Room R08 APPLIED MACRO Chair: Michael Owyang

C0203: **A quantitative analysis of the countercyclical capital buffer**
Presenter: **Miguel Faria-e-Castro**, Federal Reserve Bank of St. Louis, United States

The quantitative macroeconomic effects of the countercyclical capital buffer (CCyB) are studied in a nonlinear DSGE model with occasional financial crises, which is calibrated and combined with US data to estimate sequences of structural shocks. Raising capital buffers during leverage expansions can reduce the frequency of crises by more than half. A quantitative application to the 2007-08 financial crisis shows that the CCyB in the 2.5% range (as in the Federal Reserve's current framework) could have greatly mitigated the financial panic of 2008, for a cumulative gain of 29% in aggregate consumption. The threat of raising capital requirements is effective even if this tool is not used in equilibrium.

C0205: **Understanding growth-at-risk: A Markov-switching approach**
Presenter: **Francesca Loria**, Federal Reserve Board, United States
Co-authors: Danilo Cascaldi-Garcia, Dario Caldara, Pablo Cuba-Borda

Both financial and macroeconomic conditions matter for downside risks to the economic outlook. We show that the deterioration of the financial and real sides dramatically increase the probability of tail risks of large negative growth over the next year. We propose a real-time measure of financial conditions and economic activity for the United States, and use these measures to construct conditional quantiles and predictive distributions of average GDP growth over the next 12 months. We find that periods of high macro and financial distress, such as the Global Financial Crisis and the COVID-19 pandemic, are associated with a low average future growth, high uncertainty, and risks tilted to the downside. This methodology is a powerful tool to assess the risk of tail events, such as recessions, and to evaluate the likelihood of point forecasts.

C0206: **Back to the present: Learning about the Euro Area through a now-casting model**
Presenter: **Danilo Cascaldi-Garcia**, Federal Reserve Board, United States
Co-authors: Thiago Ferreira, Domenico Giannone, Michele Modugno

A model is built for now-casting economic conditions in the euro area and its three largest member countries—Germany, France, and Italy. The model incorporates all market moving indicators in real time. The model provides accurate predictions of economic conditions on average over the most recent 15 years, and during three historical episodes of high economic uncertainty: the Global Financial Crisis, the European sovereign debt crisis, and the onset of the Great Lockdown. Since hard data are released with a substantial delay, business sentiment surveys data provide the most informative signal of the state of the economy in real time.

C0241: **Corporate tax changes and the cost of credit**
Presenter: **Yota Deli**, UCD, Ireland

How do changes in corporate taxation affect the cost of credit? Using more than 44.000 loan deals from the US syndicated loan market from 1984 to 2017, we find that changes in the state corporate tax rates have an asymmetric effect on the cost of credit. An increase on the state corporate tax rates bears no significant results to the cost of credit, whereas a 1-point decrease of the corporate tax rate shaves at least 5.8 basis points from spreads, but likely more. The effect comes from the change of the demand of loans following a change in the corporate taxation from the firm side (demand channel), whereas the changes on the behavior of the banks (supply channel) do not bear significant results. Firm characteristics are crucial to understand the transmission mechanisms. Higher performance firms increase their demand for loans whereas firms with higher debt turn to alternative sources of financing. Our findings are robust to alternative measures of tax changes, to the comprehensive inclusion of relevant controls, to the inclusion of mechanisms for alternative sources for the financing of the firms, and to federal level controls for monetary and fiscal policy.

CC810 Room R05 CONTRIBUTIONS IN RISK ANALYSIS Chair: Simone Manganelli

C0243: **Spillover effects between commodity and stock markets: A state-dependent sensitivity expected shortfall (SDSES) approach**
Presenter: **Laura Garcia-Jorcano**, Universidad de Castilla-La Mancha, Spain
Co-authors: Lidia Sanchis

A state-dependent sensitivity expected shortfall (SDSES) model is developed which enable us to quantify the direction, size, and persistence of risk spillovers among the United States (US) and Brazil, Russia, India, China (BRIC) stock market indices and different individual commodities as a function of the state of financial markets (tranquil, normal, and volatile). For eight sets of major stock and commodity markets (SP500, BRIC stock market index, US commodity market index, oil, copper, gold, wheat, and corn), we demonstrated that spillover effects are small during normal and tranquil states and those effects are of considerable size in the volatile state and are changeable over time. We obtained high and more significant spillovers and financialization process evidence in the volatile state after Lehman Brothers bankruptcy. Market stock indices appeared to play a major role in the transmission of shocks to other markets, especially from the SP500 to wheat in the post-Lehman Brothers bankruptcy period, whereas BRIC stock market index was one of the highest sources of tail risk spillover to the oil market in the post-Draghi speech period. Furthermore, we find that oil and agricultural markets are becoming more integrated after the global financial crisis.

C0846: **Proper measures of connectedness for systemic risk detection: An application on real financial data**
Presenter: **Mishel Qyrana**, University of Pavia, Italy
Co-authors: Pierpaolo Uberti, Silvia Figini

In the context of systemic risk analysis and forecasting, some particular instances of the so-called proper measures of connectedness are imple-

mented. Proper measures of connectedness can be read as a generalization of the condition number of a matrix, allowing to consider the multiple collinearity between assets in a market. This aspect represents the loss in terms of diversification opportunities typical during systemic events. First, the choice of the parameter k, which represents the number of diversification opportunities at risk is discussed. Then, an accurate in-sample and out-of-sample application on real financial data is provided, discussing both the descriptive and predictive power of the measure. A final focus is dedicated to the out-of-sample analysis, in order to highlight when the proposed approach shows interesting forecasting power, making the measures good tools to be used as early warning signals of systemic risk.

C0179: **Loss function-based structural break detection in risk measures**
Presenter: **Xiaohan Xue**, ICMA Centre, University of Reading, United Kingdom
Co-authors: Emese Lazar, Shixuan Wang

A new test is proposed to detect change points in risk measures, based on the CUSUM procedure applied to the Wilcoxon statistics of the FZ loss function class. This efficiently captures structural breaks jointly in two risk measure series: Value-at-Risk (VaR) and Expected Shortfall (ES). We derive the asymptotic distribution of the proposed statistic. Due to the existence of nuisance parameters, we adopt a stationary bootstrapping technique to obtain the critical values of the test statistics of the loss-based Wilcoxon test. Monte Carlo simulation results justify that our proposed test has better size control and higher power than the alternative tests under various change-point scenarios. The alternatives considered include structural break detection methods based on self-normalized CUSUM statistics for the VaR series and the ES series taken individually and a modification of our proposed statistic based on Renyi-type formulation. An empirical study of the S&P 500 index illustrates that the proposed test is able to detect structural breaks in the tail of financial time series which are consistent with known market events.

C0221: **Risk as fuel of the business cycle**
Presenter: **Christoph Schult**, Halle Institute for Economic Research, Germany

The aim is to develop a dynamic stochastic general equilibrium (DSGE) model with risky capital and oil as production factors. The production function of the representative firm is a nested constant elasticity of substitution function. The model is estimated using Bayesian techniques with economic data and on oil prices, production and consumption for the United States. The interaction between risk, investment decisions of firms, and the oil market are analysed, taking the short-run elasticity of substitution between oil and capital and the propagation mechanisms between risk in capital production and oil price movements into account. The model is used to reassess the contribution of the different potential drivers to the business cycle controlling for fluctuations in oil markets. Significant findings are that the contributions of financial market frictions and oil market disturbances to the US business cycle are low and that financial market disturbances mainly drove the Great Recession. The model can quantify the impact of climate change mitigation policies on the economy. Climate change mitigation policies, e.g. increasing oil taxes, to reduce crude oil consumption by 10% can cause a contraction of GDP by 1 to 2% and increases inflation. Monetary policy can stabilize inflation increasing the federal funds rate dependent on the degree of financial market imperfections by 0.15 to 0.40 percentage points annually.

CG022 Room R04 CONTRIBUTIONS IN ECONOMETRIC INFERENCE Chair: Carsten Jentsch

C1091: **Multiple testing of the forward rate unbiasedness hypothesis across currencies**
Presenter: **Richard Luger**, Laval University, Canada
Co-authors: Hsuan Fu

Distribution-free procedures are developed to test the forward rate unbiasedness hypothesis (FRUH) across multiple currencies, jointly. These tests directly assess whether forward exchange rates provide unbiased predictions of future spot exchange rates, in levels, as predicted by the FRUH under rational expectations and risk neutrality. The approach proceeds with test statistics for individual currency FRUH assessments. It then uses Monte Carlo resampling techniques to combine the marginal p-values in a way that controls the joint significance level. Our framework allows for missing data and the presence of multivariate GARCH-type effects in the spot and forward rates. The usefulness of the new procedures is illustrated with a simulation study and with an assessment of the FRUH across 13 currencies that exist over differing time periods. We find support for the joint FRUH.

C0624: **Outlier testing in robust two-stage least squares models**
Presenter: **Jonas Kai Kurle**, University of Oxford, United Kingdom
Co-authors: Xiyu Jiao

A frequent concern in applied economics is that key empirical findings may be driven by a tiny set of outliers. To perform outlier robustness checks in instrumental variables regressions, the common practice is first to run ordinary two-stage least squares (2SLS) and classify observations with residuals beyond a chosen cut-off value as outliers. Subsequently, 2SLS is re-calculated based on the non-outlying observations, and this procedure may be iterated until robust results are obtained. However, the above trimmed 2SLS has a positive probability of finding outliers even when the data generating process contains none. To answer the question of whether observations are correctly classified as outliers, the false outlier detection rate (gauge) is studied asymptotically using an empirical processes argument. The established asymptotic theory of the gauge forms a basis for tests for the overall presence of outliers. Simulation studies lend further support to the theory, and an empirical illustration shows its utility.

C0813: **Identification-robust tests for probit models with endogenous regressors**
Presenter: **Tianyu He**, McGill University, Canada
Co-authors: Jean-Marie Dufour

Weak identification is a well-known issue in the context of linear structural models but is less studied in binary outcome models. We focus on weak identification in probit models with endogenous regressors and propose the asymptotic maximized Monte Carlo test, which is identification-robust. We compare our tests in simulation experiments to generalized minimum distance (MD) robust tests and common asymptotic tests including Wald, Lagrangian multiplier and likelihood ratio (LR) tests based on generalized method of moments (GMM), and likelihood ratio tests based on maximum likelihood estimators (MLE). We find that LR test based on MLE can have large level distortions in the presence of weak identification which is rarely documented in the literature. Meanwhile, our proposed tests control the level regardless of whether the structural parameters are identified. As for the power of tests, the simulation evidence suggests that the proposed tests exhibit reasonable power compared with MD type tests and asymptotic tests based on GMM and MLE whose critical values are locally corrected. We finally apply our method to analyze labor force participation of married female.

C0814: **Asymptotic theory and bootstrap inference for Mack's model**
Presenter: **Julia Steinmetz**, TU Dortmund University, Germany
Co-authors: Carsten Jentsch

The distribution-free chain ladder reserving model by Mack belongs to the most popular approaches in non-life insurance mathematics. As originally proposed, it serves well to determine the first two moments of the reserve distribution, but it does not allow to identify its whole distribution. To estimate also quantiles of the reserve, e.g. . to determine the value-at-risk and tail value-at-risk, Mack's model is usually equipped with a tailor-made bootstrap procedure. For this purpose, the resulting Mack bootstrap proposal requires additional parametric assumptions and postulates a normal distribution for the individual development factors. Although the Mack bootstrap is widely used in applications, no bootstrap consistency results exist that justify this approach. We establish a rigorous model framework that allows, for an increasing number of accident years, to derive asymptotic theory for the estimators in the Mack model. We investigate parametric and non-parametric implementations of the Mack bootstrap and prove bootstrap consistency results based on a suitable set of assumptions. We illustrate our findings in simulations and discuss

the validity of our approaches in situations where wrong distributional assumptions are.

CG046 Room R07 CONTRIBUTIONS IN MACHINE LEARNING TECHNIQUES IN MACROECONOMICS AND FINANCE Chair: Daniel Borup

C0204: **Talking about ESG matters**

Presenter: **Alejandro Rodriguez Gallego**, Comillas Pontifical University, Spain

Co-authors: Isabel Catalina Figuerola-Ferretti Garrigues, Sara Lumbreras Sancho

Sustainability has gained great traction among investors, however, there is no consensus in the literature about its actual impact on financial performance. The lack of quality data on ESG has hindered most attempts to tackle the issue quantitatively. The application of Natural Language Processing on a corpus of audited annual reports is proposed as a way to obtaining time series which are long and objective enough for empirical methods. Taking a sample of more than 3,100 companies, high versus low portfolios are built based on the new scores for Environment, Social and Governance and new Fama-French factors are derived from them. The results show that talking about Social and Governance has a moderating effect over volatility while increasing risk-adjusted return. Conversely, the effect reverses for Environment. Finally, the loading coefficients estimated with panel data suggest that industries interpret sustainability discourse differently.

C0643: **Financial conditions and economic activity: Insights from machine learning**

Presenter: **Michael Kiley**, Federal Reserve Board, United States

Machine learning (ML) techniques are used to construct a financial conditions index (FCI). The components of the ML-FCI are selected based on their ability to predict the unemployment rate one-year ahead. Three lessons for macroeconomics and variable selection/dimension reduction with large datasets emerge. First, variable transformations can drive results, emphasizing the need for transparency in the selection of transformations and robustness to a range of reasonable choices. Second, there is strong evidence of nonlinearity in the relationship between financial variables and economic activity: tight financial conditions are associated with sharp deteriorations in economic activity, and accommodative conditions are associated with only modest improvements in activity. Finally, the ML-FCI places sizable weight on equity prices and term spreads, in contrast to other measures. These lessons yield an ML-FCI showing tightening in financial conditions before the early 1990s and early 2000s recessions, in contrast to the National Financial Conditions Index (NFCI).

C0936: **Multilingual entity resolution with semi-supervised machine learning techniques: An application to KYC**

Presenter: **Ilgiz Mustafin**, Innopolis University, Russia

Co-authors: George Emelyanov, Marius Frunza

The aim is to explore several entity resolution strategies that can be used when the various datasets are recorded in different languages. When trying to link records from different datasets we encounter two main issues. On the one hand, matching records names in different languages brings a significant challenge due to the fact that transliteration is not a bijective well-defined function. On the other hand, matching records based solely on names generates a high number of false-positives. For the first issue, we apply a strategy based on neural networks trained on a predefined dataset on which we add a set of expert-based rules. We address the second issue with a Bayesian approach tacking into account the apriori distribution of names frequencies in our datasets. Our approach is tested on business registries datasets extracted from different sources like government registers or leaked documents. The most notable difference is the way how names are presented in different languages of the documents. The existing techniques are extended, and a method is presented for finding matches in business networks effectively. The results of this research can be used for coupling national business registers containing local data about companies and owners recorded in multiple languages. Thus, the suggested approach can improve the frameworks used currently in the Know Your Client (KYC) process.

C0768: **Measuring the impact of President Trump's tweets on economic uncertainty: A narrative approach**

Presenter: **Hector Daniel Perico Ortiz**, Friedrich-Alexander-Universitat Erlangen-Nurnberg, Germany

The causal relation between President Trump's tweeting behavior and market uncertainty at a high-frequency level is investigated. We implement an economic narrative approach based on identification of economic narratives from Twitter data using machine learning and estimating the effect of them, using time series regressions, on a high frequency measure of market uncertainty, given by the five-minute change in the VIX index. The results suggest that major economic narratives regarding foreign trade and policy, and monetary policy, have a significant effect on market uncertainty in the period of one hour and three hours after the narrative event. Immigration narrative is also significant at the five hours horizon. Furthermore, behavior events, such as increases in the tweet or retweeted counts above their average, matter for market uncertainty. A similar analysis at the daily frequency level using the EPU index as uncertainty measure provides similar results at longer time horizons.

Monday 21.12.2020 16:05 - 17:45 Parallel Session Q – CFE-CMStatistics

EI013 Room R14 STATISTICAL FOUNDATION FOR DATA SCIENCE **Chair: Soutir Bandyopadhyay**

E0158: **Consistent variational Bayes neural networks classification with application to an Alzheimer disease study**
Presenter: **Shrijita Bhattacharya**, Michigan State University, United States
Co-authors: Zihuan Liu, Taps Maiti

Bayesian neural networks models (BNN) have re-surged recently due to the advancement of scalable computations and its utility in solving complex prediction problems in applications such as medical image analysis and computer vision tasks. However, the conventional Markov Chain Monte Carlo (MCMC) based implementation suffers from various issues such as computational costs, finding suitable proposal densities, etc. which limit the use of this powerful technique in large scale studies. The variational Bayesinference has become a viable alternative to circumvent some of the computational issues. Although the approach is popular in machine learning, its application in statistics is somewhat limited. A variational BNN estimation methodology and related tactical theory are developed. The numerical algorithms and their practical aspects are discussed in detail. The theory for posterior consistency, a desirable property in nonparametric Bayesian statistics, is also developed. The theory provides an assessment of prediction accuracy and guidelines for characterizing the prior distributions and variational family. The loss of using a variational posterior over the true posterior has also been quantified. The development is motivated by an important application in biomedical engineering, namely building predictive tools for the transition from mild cognitive impairment (MCI) to Alzheimer disease (AD) and emphasizing clinical aspects of the field.

E0273: **Consistent sparse deep learning: Theory and computation**
Presenter: **Faming Liang**, Purdue University, United States

Deep learning has been the engine powering many successes of data science. However, the deep neural network (DNN), as the basic model of deep learning, is often excessively over-parameterized, causing many difficulties in training, prediction and interpretation. We propose a frequentist-like method for learning sparse DNNs and justify its consistency under the Bayesian framework: the proposed method could learn a sparse DNN with at most $O(n/\log(n))$ connections and nice theoretical guarantees such as posterior consistency, variable selection consistency and asymptotically optimal generalization bounds. In particular, we establish posterior consistency for the sparse DNN with a mixture Gaussian prior, show that the structure of the sparse DNN can be consistently determined using a Laplace approximation-based marginal posterior inclusion probability approach, and use Bayesian evidence to elicit sparse DNNs learned by an optimization method such as stochastic gradient descent in multiple runs with different initializations. The proposed method is computationally more efficient than standard Bayesian methods for large-scale sparse DNNs. The numerical results indicate that the proposed method can perform very well in large-scale network compression as well as feature selection for high-dimensional nonlinear regression, both advancing interpretable machine learning.

E0155: **The dependable data-driven discovery institute**
Presenter: **Dan Nettleton**, Iowa State University, United States

The Departments of Computer Science, Mathematics, and Statistics at Iowa State University are collaborating to form the Dependable Data-Driven Discovery (D4) Institute with the support of a United States National Science Foundation TRIPODS (Transdisciplinary Research in Principles of Data Science) grant. A brief overview of the D4 Institute and associated research activities will be provided, with a focus on one D4 thrust that involves quantifying prediction uncertainty in COVID-19 infection forecasts, agricultural applications, and sports analytics.

EO315 Room R04 STATISTICAL MODELS: RECENT DEVELOPMENTS II **Chair: Anna Panorska**

E1160: **Simulating high-dimensional multivariate data using the bigsimr R package**
Presenter: **Alfred Schissler**, University of Nevada, Reno, United States
Co-authors: Anna Panorska, Tomasz Kozubowski, Alex Knudson, Juli Petereit

It is critical to simulate data when conducting Monte Carlo studies and methods realistically. But measurements are often correlated and high dimensional in this era of big data, such as data obtained through high-throughput biomedical experiments. Due to computational complexity and a lack of user-friendly software available to simulate these massive multivariate constructions, researchers often resort to simulation designs that posit independence. This greatly diminishes insights into the empirical operating characteristics of any proposed methodology, such as false-positive rates, statistical power, interval coverage, and robustness. This talk introduces the bigsimr R package that provides a general, scalable procedure to simulate high-dimensional random vectors with given marginal characteristics and dependency measures. We'll describe the functions included in the package, including multi-core and graphical-processing-unit accelerated algorithms to simulate random vectors, estimate correlation matrices, and find close positive semi-definite matrices. Finally, we showcase the power of bigsimr by applying these functions to our motivating dataset — RNA-sequencing data obtained from breast cancer tumor samples with sample size $n = 1212$ patients and dimension $d > 1000$.

E1163: **An algorithmic method for fitting multimodes heavy-tailed distributions**
Presenter: **Marie Kratz**, ESSEC Business School, CREAR, France

A hybrid model is proposed for heavy-tailed phenomena, combining a Gaussian Mixture Model (GMM) with a Generalized Pareto component (GPD). It generalizes a previous algorithmic method, with an automatic detection of the tail threshold. While the main advantage of the GMM is the flexibility it affords when dealing with multimodal distribution, introducing a GPD component allows one to evaluate the tail of the distribution, unlike a pure GMM.

E1170: **A generalized linear model for extreme events: New goodness of fit measures**
Presenter: **Francesco Zuniga**, University of Nevada at Reno, United States
Co-authors: Anna Panorska, Tomasz Kozubowski

A generalized linear model is presented for the observed vector (N,X,Y) of duration N, magnitude X and maximum Y of an (often extreme) event such as flood, extreme precipitation, market growth or decline. Out GLM allows for the investigation of the association of the (N,X,Y) with covariates. We also present new approach of goodness of fit assessment. The methodology is illustrated by modeling growth events of the S&P500 index on covariates such as unemployment and inflation rate.

E1177: **Branching out: New techniques in tree representations of time series with applications**
Presenter: **Zoe Haskell**, University of Nevada, Reno, United States
Co-authors: Ilya Zaliapin

Tree representation of continuous functions has historically been used to explore the properties of trees and stochastic processes. However, its usage in an applied analysis of time series remains limited. We introduce new objects, partial trees and partial Harris paths, that facilitate mapping between finite time series and rooted plane trees, and establish correspondences between natural operations on trees and their time-series counterparts. We show that the Horton pruning of a tree (cutting leaves and subsequent series reduction) corresponds to taking the local minima of a time series. We also present new pruning algorithms adopted for the analysis of spiky time series and of series with significant flat periods (plateaus). We illustrate the proposed techniques in the problems of noise reduction via pruning in tree-space, and detecting clusters in sets of time series data, including work done at Disney for audience segmentation.

EO658 Room R05 EFFICIENT NONPARAMETRIC METHODS FOR COMPLEX DATA Chair: Daniel McDonald

E0554: **Dyadic CART revisited**
Presenter: **Sabyasachi Chatterjee**, University of Illinois at Urbana Champaign, United States
The algorithm Dyadic CART for nonparametric regression is revisited. We explore methods related to Dyadic CART in the context of some nonparametric function estimation problems of recent interest.

E0953: **Efficient estimation of smooth functionals in Gaussian shift models**
Presenter: **Mayya Zhilova**, Georgia Institute of Technology, United States
Co-authors: Vladimir Koltchinskii
A problem of estimation of smooth functionals of a high-dimensional parameter of a Gaussian shift model with known covariance operator is studied. We develop asymptotically efficient estimators for functionals of Hoelder smoothness s. We show that their mean squared error rate is minimax optimal, at least in the case of finite-dimensional standard Gaussian shift model. Moreover, we determine a sharp threshold on the smoothness s such that for all s above the threshold, the functional can be estimated efficiently in a "small noise" setting. The construction of the efficient estimators is crucially based on a bootstrap chain method of bias reduction.

E0961: **Quantile trend filtering**
Presenter: **Oscar Hernan Padilla**, UCLA, United States
Quantile trend filtering, a recently proposed method for one-dimensional nonparametric quantile regression, is studied. We show that the penalized version of quantile trend filtering attains minimax rates, off by a logarithmic factor, for estimating the vector of quantiles when its kth discrete derivative belongs to the class of bounded variation signals. Our results also show that the constrained version of trend filtering attains minimax rates in the same class of signals. Furthermore, we show that if the true vector of quantiles is piecewise polynomial, then the constrained estimator attains optimal rates up to a logarithmic factor. We also illustrate how our technical arguments can be used for analyzing other shape constrained problems with quantile loss. Finally, we provide extensive experiments that show that quantile trend filtering can perform well, based on mean squared error criteria, under Cauchy and other heavy-tailed distributions of the errors.

E1084: **Trend filtering on regular lattices with sub-exponential noise**
Presenter: **Veeranjaneyulu Sadhanala**, University of Chicago, United States
Co-authors: Daniel McDonald, Robert Bassett, James Sharpnack
Trend filtering is a locally adaptive nonparametric regression method that fits a piecewise polynomial to univariate data with automatically chosen knots. The statistical performance of trend filtering and its extensions to regular lattices was analyzed recently, assuming that the observations are corrupted by sub-gaussian noise. We study this problem with sub-exponential noise. We derive minimax optimal error bounds on mean which are same as in the sub-gaussian case but with an additional logarithmic factor. We also argue why it is hard to upper bound the KL divergence, which is a more natural quantity to bound in the exponential family setting.

EO698 Room R11 FUNCTIONAL AND COMPLEX DATA ANALYSIS Chair: Paromita Dubey

E0331: **Functional autoregressive processes via reproducing kernel Hilbert spaces**
Presenter: **Daren Wang**, University of Chicago, China
The aim is to study the estimation and prediction of a functional autoregressive (FAR) process, a statistical tool for modeling functional time series data. Due to the infinite-dimensional nature of FAR processes, the existing literature addresses its inference via dimension reduction and theoretical results therein require the assumption of fully observed functional time series. We propose an alternative inference framework via the tools of Reproducing Kernel Hilbert Spaces (RKHS). Specifically, a nuclear norm regularization method is proposed for estimating the transition operators of the FAR process directly from discrete samples of the functional time series. We derive a Representer theorem for the FAR process, which enables infinite-dimensional inference without dimension reduction. Consistent theoretical guarantees are established under the (more realistic) assumption that we only have finite discrete samples of the FAR process.

E0355: **Functional models for time-varying random objects**
Presenter: **Hans-Georg Mueller**, University of California Davis, United States
Co-authors: Paromita Dubey
In recent years, samples of time-varying object data such as time-varying networks that are not in a vector space have become increasingly prevalent. Such data are elements of a general metric space that lacks local or global linear structure. Common approaches that have been used with great success for the analysis of functional data, such as functional principal component analysis, are therefore not applicable. The concept of metric covariance makes it possible to define a metric auto-covariance function for a sample of random curves that take values in a general metric space. This metric auto-covariance function is non-negative definite when the squared semi-metric of the underlying space is of negative type. Then the eigenfunctions of the linear operator with the auto-covariance function as kernel can be used as building blocks for an object functional principal component analysis, which includes real-valued Frechet scores and metric-space valued object functional principal components. Sample-based estimates of these quantities are shown to be asymptotically consistent and are illustrated with time-varying networks and other data.

E0925: **Geometric approaches to inference: Non-Euclidean data and networks**
Presenter: **Dena Asta**, The Ohio State University, United States
The purpose is to describe applications of geometry to large-scale data analysis. An overriding theme is that an understanding of the relevant geometric structure in the data is useful for efficient and large-scale statistical analyses. Firstly, we will discuss geometric methods for non-parametric methods in non-Euclidean spaces. Secondly, we will discuss a geometric approach to network inference by focusing on the Riemannian geometry of CLS models.

E1148: **Frechet random forests for metric space valued regression with non-Euclidean predictors**
Presenter: **Louis Capitaine**, Bordeaux University INSERM, France
Random forests are a statistical learning method widely used in many areas of scientific research because of its ability to learn complex relationships between input and output variables and also their capacity to handle high-dimensional data. However, current random forest approaches are not flexible enough to handle heterogeneous data such as curves, images and shapes. We introduce Frechet trees and Frechet random forests, which allow handling data for which input and output variables take values in general metric spaces (which can be unordered). To this end, a new way of splitting the nodes of trees is introduced, and the prediction procedures of trees and forests are generalized. Then, random forests out-of-bag error and variable importance score are naturally adapted. A consistency theorem for Frechet regressogram predictor using data-driven partitions is given and applied to Frechet purely uniformly random trees. The method is studied through several simulation scenarios on heterogeneous data combining longitudinal, image and scalar data. Finally, a dataset from an HIV vaccine trial is analyzed with the proposed method.

EO243 Room R12 ADVANCES IN THE ANALYSIS OF FUNCTIONAL DATA AND FUNCTIONAL TIME SERIES Chair: Anne van Delft

E0409: **Two-sample tests for relevant differences in the eigenfunctions of covariance operators**
Presenter: **Alexander Aue**, UC Davis, United States

Co-authors: Holger Dette, Gregory Rice

Two-sample tests for functional time series data are considered. These tests have become widely available in conjunction with the advent of modern complex observation systems. Particular interest is in finding out whether two sets of independent functional time-series observations share the shape of their primary modes of variation as encoded by the eigenfunctions of the respective covariance operators. To this end, a novel testing approach is introduced that connects with, and extends, existing literature in two main ways. First, tests are set up in the relevant testing framework, where interest is not in testing an exact null hypothesis but rather in detecting deviations deemed sufficiently significant, with significance often determined in consultation with scientific guidelines. Second, the proposed test statistics rely on a self-normalization principle that helps avoid the notoriously difficult task of estimating the long-run covariance structure of the underlying functional time series. The main theoretical result to be discussed is the derivation of the large-sample behavior of the proposed test statistics. Empirical evidence, indicating that the proposed procedures work well in finite samples, is provided through a simulation study, also comparing with competing methods and an application to annual temperature data.

E0531: **Preprocessing functional data by a factor model approach**

Presenter: **Siegfried Hoermann**, Graz University of Technology, Austria

Co-authors: Fatima Jammoul

Functional data measured on a discrete set of observation points are considered. Often such data are measured with noise, and then the target is to recover the underlying signal. Commonly this is done with some smoothing approach, e.g. kernel smoothing or spline fitting. While such methods act function by function, we argue that it is more accurate to take into account the entire sample for the data preprocessing. To this end we propose to fit factor models to the raw data. We show that the common component of the factor model corresponds to the signal which we are interested in, whereas the idiosyncratic component is the noise. Under mild technical assumptions we demonstrate that our estimation scheme is uniformly consistent. From a theoretical standpoint our approach is elegant, because it is not based on smoothness assumptions and generally permits a realistic framework. The practical implementation is easy because we can resort to existing tools for factor models. OurThe empirical investigations provide convincing results.

E0648: **Functional GARCH-X Model with an application to forecasting crude oil return curves**

Presenter: **Yuqian Zhao**, University of Essex, United Kingdom

Co-authors: Gregory Rice, Tony Wirjanto

Functional data objects derived from high-frequency financial data are uncorrelated but long-range conditionally heteroscedastic. The existing functional GARCH models are designed to account for conditional heteroscedasticity, but not specifically to capture long-range dependent dynamics in the data. We propose a functional GARCH-X model, where the covariate X is chosen to be weakly stationary with a long-range dependence property. The functional autocorrelation coefficients of the squared process of this and other recently introduced functional volatility processes are studied. Monte Carlo simulation shows that the functional autocorrelation coefficients of the squared functional GARCH-X process behave closely to those observed in the empirical data. In an empirical application, we forecast conditional volatility of the WTI crude oil intra-day return curves collected from the commodity futures market. The results show that the FGARCH-X model provides mild corrections to the functional volatility models in terms of the conditional volatility prediction, yielding more precise confidence bands for the return curves.

E0721: **Testing for the rank of a covariance operator**

Presenter: **Anirvan Chakraborty**, IISER Kolkata, India

Co-authors: Victor Panaretos

How we can discern whether the covariance operator of a stochastic process is of reduced rank. If so, what is its rank, and can we say so at a given confidence level? This question is central to several problems in functional data, which require low-dimensional representations. The difficulty is that the determination has to be made based on discrete observations with measurement errors. This adds a ridge to the empirical covariance, obfuscating the underlying dimension. We discuss a matrix-completion inspired test that circumvents this issue by measuring the optimum least-square fit of the empirical covariance's off-diagonal elements, over finite rank covariances. For a sufficiently large but fixed grid, we discuss the asymptotic null distribution as the sample size grows. We use it to construct a bootstrap implementation of a stepwise testing procedure controlling the FWER corresponding to the collection of hypotheses formalising the problem. Under certain regularity assumptions, we show that the procedure is consistent and its bootstrap implementation is valid. The procedure circumvents smoothing, is indifferent to hetreroskedastic errors, and does not assume a low-noise regime. We show the excellent practical performance on simulations and real data, and also demonstrate the stability across a wide range of settings.

EO570 Room R13 ADVANCES IN THE ANALYSIS OF LARGE AND COMPLEX DATA Chair: Jing Qiu

E0705: **Asymmetric Laplace distribution jumps in continuous-time financial modelling**

Presenter: **Matthew Stuart**, Iowa State University, United States

Co-authors: Cindy Yu

In the existing continuous-time finance literature with jumps in returns and stochastic volatility (SV), it is often assumed that the return jumps are normally distributed with a negative mean, the volatility jumps are exponentially distributed, and the jumps occur either contemporaneously or independently, but not both. We propose to use an asymmetric Laplace distribution (ALD) to model jumps in returns and volatility (contemporaneous or independent) in order to overcome the drawback of lack of monotonicity in jump size density due to using a normal distribution with a negative mean. We also further the new research into cryptocurrency markets by proposing an ALD in returns in a 2-dimensional dataset, specifically on a market index and cryptocurrency, to examine the relationship between the assets. Monte Carlo Markov Chain (MCMC) methods are developed to estimate the model parameters and latent state variables, such as SV, jump times, jump sizes, and is validated through simulation studies. The method is applied to fit both the S&P 500 and Bitcoin independent and joint daily returns from 2014 to 2020.

E0738: **Model-based detection of differential causality between large networks**

Presenter: **Dabao Zhang**, Purdue University, United States

A novel statistical method is developed to identify causal differences between two cohorts characterized by structural equation models. We propose to reparameterize the model to separate the differential structures from common structures, and then design an algorithm with calibration and construction stages to identify these differential structures. The calibration stage serves to obtain consistent prediction by building the L_2 regularized regression of each endogenous variables against pre-screened exogenous variables, correcting for potential endogeneity issue. The construction stage consistently selects and estimates both common and differential effects by undertaking L_1 regularized regression of each endogenous variable against predicts of other endogenous variables as well as its anchoring exogenous variables. Our method allows easy parallel computation at each stage. Theoretical results are obtained to establish non-asymptotic error bounds of predictions and estimates at both stages, as well as the consistency of identified common and differential effects. The studies on synthetic data demonstrated that the proposed method performed much better than independently constructing the networks. A real data set is analyzed to illustrate the applicability of our method.

E0914: **A Bayesian approach to identify genes with multiple expression patterns for paired RNA-seq data**

Presenter: **Jing Qiu**, University of Delaware, United States

Co-authors: Zhuoqiong He, Yuanyuan Bian

It is often of interest to identify genes with specific expression patterns over several conditions such as time points, genotypes, etc. The common practice is to perform differential expression analysis separately for each condition and then combine the results to obtain a list of genes with desired expression pattern or profiles. Such practice can inflate the type I error for identifying genes with different expression patterns under multiple conditions, especially when the desired expression pattern involves equally expression under certain conditions. We propose a Bayesian approach to identify genes with multiple expression patterns under two conditions with FDR controlled for all desired expression patterns simultaneously. The inverse moment non-local prior is used for modeling expression patterns with equal expression under one condition. Our simulation studies show that it is a much more challenging job to identify genes that are equally expressed in one condition but differentially expressed in the other condition than identify genes that are differentially expressed in both conditions. The common practice in literature can have highly inflated type I error for identifying the former type of genes. Our method has FDR controlled close to the nominal level with better power than the popular methods in the literature.

E1056: **Set testing methods based on zero-inflated models for microbiome data**
Presenter: **Chong Wang**, Iowa State University, United States

With advances in sequencing methods, the study of the microbiome has greatly increased. Microbiome data, in the form of an OTU or ASV count table, can be used to identify specific ASVs that function differently across treatment conditions. Such analysis is deemed differential abundance analysis. ASVs are grouped by their taxonomic rank, and ASVs sharing the same rank have similar biological traits. By studying groups or sets of ASVs, and identifying if the set is differentially abundant, the biological interpretation of a microbiome study is enhanced. We review current approaches in set testing methods and apply them to a microbiome data set from a 2017 study. We propose a new set testing method based on an existing Poisson hurdle model and compare performance across all methods through a simulation study. We find that our proposed model outperforms existing approaches with zero-inflated observations.

EO516 Room R15 COMPUTATIONAL ISSUES IN INFECTIOUS DISEASE EPIDEMIOLOGY Chair: Rob Deardon

E0231: **Spatio-temporal Bayesian modeling of county level Covid-19 in South Carolina**
Presenter: **Andrew Lawson**, Medical University of South Carolina, United States
Co-authors: Joanne Kim

Covid-19 has spread around the world and has become a pandemic in 2020. Locally within the US, the spread of the disease had been highly variable and considerable spatial heterogeneity has been apparent. In addition, data quality issues abound. We outline a spatially-referenced susceptible-infected-removed (SIR) model that can be used to describe the dynamics of symptomatic transmission. We also treat asymptomatic cases as latent variables. Deaths are also modeled. The modeling is at county level in South Carolina, although other spatial scales could be examined using these tools. Prediction of outbreaks and general future events is also discussed.

E0488: **Approximate Bayesian Computation and history matching for inference in infectious disease systems**
Presenter: **Trevelyan McKinley**, University of Exeter, United Kingdom

Complex mathematical models are being increasingly used to inform decision making. Adequately capturing key sources of uncertainty is important to produce robust predictions and reduce the probability of making poor decisions. Approximate Bayesian Computation (ABC) and other simulation-based inference methods are becoming increasingly popular for inference in complex systems, particularly ones where the likelihood function is intractable. This is due to their relative ease-of-implementation compared to alternative approaches, since they require only the means to simulate from the underlying complex model. However, despite their utility, scaling simulation-based methods to fit large-scale systems introduces a series of additional challenges that hamper robust inference. Here we use a real-world model of HIV transmission - that has been used to explore the potential impacts of potential control policies in Uganda - to illustrate some of these key challenges when applying ABC methods to high dimensional, computationally intensive models. We then discuss an alternative approach - history matching with emulation - that aims to address some of these issues and conclude with a comparison between these different methodologies.

E0432: **Machine learning assisted infectious disease modelling**
Presenter: **Rob Deardon**, University of Calgary, Canada

Information obtained from statistical infectious disease transmission models can be used to inform the development of containment strategies. Inference procedures such as Bayesian Markov chain Monte Carlo are typically used to estimate parameters of such models, but are often computationally expensive. However, in an emerging epidemic, stakeholders must move quickly to contain spread. We explore machine learning methods for carrying out fast inference via supervised classification. We consider the issues of regularization, model choice and parameter estimation. This is done within the context of spatial models, applied to both diseases of agriculture and the COVID-19 epidemic. We also consider how accurate such methods are in comparison with nave, and much slower, MCMC approaches.

E0823: **Complex models, time pressure, and COVID**
Presenter: **Thomas House**, University of Manchester, United Kingdom

The focus is on some of the issues that arose while providing modelling and statistical support to the UK Government during the current pandemic. Given the fast-changing nature of the pandemic, the exact content is subject to change; however, questions include (1) Use of parametric versus non-parametric versus mechanistic models; (2) When to use asymptotic approximations (e.g. Laplace) that may be optimistic about uncertainty but allow for much faster calculations; (3) When to combine multiple analyses sources formally and when to run separate analyses and compare using expert judgement. These are, of course, questions of general interest in computational statistics.

EO123 Room R16 METHODS FOR MISSING DATA IN EHR-BASED STUDIES Chair: Sebastien Haneuse

E0635: **Robust causal inference for point exposures with missing confounders**
Presenter: **Alexander Levis**, Harvard T.H. Chan School of Public Health, United States
Co-authors: Sebastien Haneuse

The gold standard inferential target in comparative effectiveness research is a causal treatment effect. When clinical trials are infeasible, observational cohorts can be used to estimate these effects, but statistical methods are needed to control for confounding factors. Such cohorts, especially when extracted from large observational databases, are often subject to large amounts of missing data, and methods must simultaneously handle confounding and missingness. We propose robust, semiparametric efficient estimators of average treatment effects from cohort studies when confounders are missing at random. The approach is based on a novel factorization of the likelihood that, unlike alternative methods, facilitates flexible modelling of nuisance functions while still maintaining consistency at nominal rates of convergence. Simulations, derived from an EHR-based study of the long-term effects of bariatric surgery on weight outcomes, verify the robustness properties of the proposed estimators in finite samples. Extensions of the methods to the matched cohort study design are discussed.

E0704: **Sensitivity analysis for missing not at random data in electronic health records-based research**
Presenter: **Tanayott Thaweethai**, Massachusetts General Hospital, United States
Co-authors: Sebastien Haneuse

While electronic health records (EHR) present a rich and promising data source for conducting observational research, they are highly susceptible to missingness due to the complex process by which EHR are collected and generated. Even worse, data in EHR is frequently missing not at random

(MNAR); e.g., whether a given laboratory test is ordered is often correlated with the expected value of the test. Building off a novel framework for handling missing data in EHR based on a modularization of the data provenance (i.e., the process by which data is observed in EHR), we present a method for localizing sensitivity to MNAR data to specific decisions or actions made by patients, their healthcare providers, or the larger healthcare system. We conclude with novel strategies for interpreting the results of multidimensional sensitivity analyses for MNAR data.

E0949: **A Bayesian nonparametric approach for missing data for causal inference in EHRs that uses auxiliary information**
Presenter: **Michael Daniels**, University of Florida, United States
Co-authors: Sebastien Haneuse, David Lindberg

A Bayesian nonparametric using enriched Dirichlet process mixtures is proposed to model the observed data in EHRs with an ultimate goal of causal inference. Missing data (in confounders and the outcome) is allowed to be nonignorable and auxiliary information in the EHR can be exploited to move the missingness closer to MAR. We illustrate the approach via simulations and a data example.

E1159: **Missing data and multiple imputation in clinical epidemiological research**
Presenter: **Irene Petersen**, UCL, United Kingdom

Missing data are ubiquitous in clinical epidemiological research and often found in electronic health records. Individuals with missing data may differ from those with no missing data in terms of the outcome of interest and prognosis in general. Missing data can constitute considerable challenges in the analyses and interpretation of results and can potentially weaken the validity of results and conclusions. Several ad-hoc methods have been developed for dealing with missing data. These include complete-case analyses, missing indicator method, single value imputation, and sensitivity analyses incorporating worst-case and best-case scenarios. If applied under the missing completely at random (MCAR) assumption, some of these methods can provide unbiased but often less precise estimates. Multiple imputation is an alternative method to deal with missing data, which accounts for the uncertainty associated with missing data and provides unbiased and valid estimates of associations based on information from the available data. We will discuss the different methods for dealing with missing data in clinical epidemiology.

EO435 Room R17 ADVANCES AND CHALLENGES IN MICROBIOME DATA ANALYSES Chair: Zhigen Zhao

E0391: **Rarefaction-based extensions of the LDM and PERMANOVA for testing presence-absence associations in the microbiome**
Presenter: **Yijuan Hu**, Emory University, United States
Co-authors: Andrea Lane, Glen Satten

Many methods for testing association between the microbiome and covariates of interest assume that these associations are driven by changes in the relative abundance of taxa. However, these associations may also result from changes in which taxa are present and which are absent. Analyses of such presence-absence associations face a unique challenge: confounding by library size. It is known that rarefaction controls this bias, but at the potential cost of information loss as well as the introduction of a stochastic component in the analysis. Currently, there is a need for robust and efficient methods for testing presence-absence associations in the presence of such confounding, both at the community level and at the individual-taxon level, that avoid the drawbacks of rarefaction. Here we present extensions of the LDM and PERMANOVA for testing presence-absence associations. The extended LDM and PERMANOVA are both non-stochastic approaches that repeatedly apply the LDM and PERMANOVA to all rarefied taxa count tables, averages the residual sum-of-squares (RSS) terms over the rarefaction replicates, and then forms an F-statistic based on these average RSS terms. Our simulations indicate that our proposed methods are robust to any systematic differences in library size and have better power than alternative approaches. We illustrate our method using an analysis of data on inflammatory bowel disease (IBD) in which cases have systematically smaller library sizes than controls.

E0718: **Compositional mediation models: Application to microbiome data**
Presenter: **Michael Sohn**, University of Rochester, United States

The importance of the microbiome in maintaining human health and its contribution to disease when it is perturbed has been well established. It is also well known that the microbiome is shaped by external factors, such as diet and medication. Therefore, understanding the mediating role of the microbiome in linking external factors and our health conditions is crucial to translate the microbiome research into therapeutic and preventative applications. We introduce sparse compositional mediation models under potential outcomes framework to estimate and test the causal mediation effect (i.e., mediation effects of the microbiome) utilizing the compositional algebra defined in the simplex space and a linear zero-sum constraint on regression parameters.

E0393: **ConQuR: Batch effect correction for microbiome data via conditional quantile regression**
Presenter: **Wodan Ling**, Fred Hutchinson Cancer Research Center, United States
Co-authors: Michael C Wu

Mega-analysis by integrating batches of data boosts the power to detect associations between microbiome data and clinical variables of interest. However, as with other high-throughput data, microbiome data can suffer from severe batch effects, which simultaneously leads to excessive false positives and false negatives. Most of the existing strategies for mitigating batch effects in microbiome data rely on approaches originally designed for genomic analysis. Many of them assume Gaussian linear or negative binomial regression models, which fail to adequately address the severe zero-inflation, dispersion and heterogeneity issues in microbiome data. The other strategies tailored for microbiome data can only be used for association testing, which fails to allow other common analytic goals such as visualization. Moreover, some of them require particular types of controls/spike-ins, making them not applicable to different designs. We developed ConQuR, a batch correction method, which uses a two-part quantile regression model to consider both inflated zeros and complex distributional attributes of the non-zero measures. It preserves the zero-inflated integer nature of microbiome data, which is compatible with any subsequent microbiome normalization and analysis. We applied ConQuR to several real data sets and showed that it outperforms the existing methods in removing batch effects and boosting the power to detect associations from the data.

E0564: **IFAA: Robust association identification and inference for absolute abundance in microbiome analyses**
Presenter: **Zhigang Li**, Department of Biostatistics, University of Florida, United States

The target of inference in microbiome analyses is usually relative abundance (RA) because RA in a sample (e.g., stool) can be considered as an approximation of RA in an entire ecosystem (e.g., gut). However, inference on RA suffers from the fact that RA is calculated by dividing absolute abundances (AA) over the common denominator (CD), the summation of all AA (i.e., library size). Because of that, a perturbation in one taxon will result in a change in the CD and thus cause false changes in RA of all other taxa, and those false changes could lead to false-positive/negative findings. We propose a novel analysis approach (IFAA) to make robust inference on AA of an ecosystem that can circumvent the issues induced by the CD problem and compositional structure of RA. IFAA can also address the issues of overdispersion and handle zero-inflated data structures. IFAA identifies microbial taxa associated with the covariates in Phase one and estimates the association parameters by employing an independent reference taxon in Phase two. Two real data applications are presented, and extensive simulations show that IFAA outperforms other established existing approaches by a big margin in the presence of unbalanced library size.

EO728 Room R18 STATISTICAL INFERENCE IN COMPLEX NETWORKS Chair: Subhadeep Paul

E0872: **Non-uniform sampling of fixed margin binary matrices**
Presenter: **Bailey Fosdick**, Colorado State University, United States

Co-authors: Alex Fout

Data sets in the form of binary matrices are ubiquitous across scientific domains, and researchers are often interested in identifying and quantifying noteworthy structure. One approach is to compare the observed data to that which might be obtained under a null model. We consider sampling from the space of binary matrices which satisfy a set of marginal row and column sums. Whereas existing sampling methods have focused on uniform sampling from this space, we introduce modified versions of two elementwise swapping algorithms which sample according to a non-uniform probability distribution defined by a weight matrix, which gives the relative probability of a one for each entry. We demonstrate that values of zero in the weight matrix, i.e. structural zeros, are generally problematic for swapping algorithms, except when they have special monotonic structure. We explore the properties of our algorithms through simulation studies. We also illustrate the potential impact of employing a non-uniform null model using a classic bird habitation dataset.

E0235: **Testing correlation of unlabeled random graphs**

Presenter: **Jiaming Xu**, Duke University, United States

Co-authors: Yihong Wu, Sophie Yu

The problem of detecting the edge correlation between two random graphs with unlabeled nodes is studied. This is formalized as a hypothesis testing problem, where under the null hypothesis, the two graphs are independently generated; under the alternative, the two graphs are edge-correlated under some latent node correspondence, but have the same marginal distributions as the null. For both Gaussian-weighted complete graphs and dense ER graphs, we determine the sharp threshold at which the optimal testing error probability exhibits a phase transition from zero to one. For sparse ER graphs, we determine the threshold within a constant factor. The proof of the impossibility results is an application of the conditional second-moment method, where we bound the truncated second moment of the likelihood ratio by carefully conditioning on the typical behavior of the intersection graph (consisting of edges in both observed graphs) and taking into account the cycle structure of the induced random permutation on the edges.

E0359: **A nonparametric test of co-spectrality in networks**

Presenter: **Srijan Sengupta**, North Carolina State University, United States

Living in an interconnected world where network-valued data arises in many domains, statistical network analysis has emerged as an active area. However, the topic of hypothesis testing in networks has received relatively less attention. We consider the problem where one is given two networks, and the goal is to test whether the given networks are cospectral, i.e., they have the same non-zero eigenvalues. Cospectral graphs have been well studied in graph theory and computer science. Cospectrality is relevant in real-world networks since it implies that the two networks share several important path-based properties, such as the same number of closed walks of any given length, the same epidemic threshold, etc. However, to the extent of our knowledge, there has not been any formal statistical inference work on this topic. We propose a non-parametric test of co-spectrality by leveraging some recent developments in random matrix theory. We develop two versions of the test — one based on an asymptotic bound and one based on bootstrap resampling. We establish theoretical results for the proposed test and demonstrate its empirical accuracy using synthetic networks sampled from a wide variety of models as well as several well-known real-world network datasets.

E1141: **Posterior predictive differences for comparing multiple networks**

Presenter: **Anna Smith**, University of Kentucky, United States

Co-authors: Tian Zheng

Comparing multiple networks, in a way that appropriately addresses nodal dependence, is a difficult task. It is complicated by the fact that many networks under comparison vary in size, the number of nodes in the network. We consider Bayesian versions of latent space models for network data, which model nodes who are closer together in a (typically Euclidean) latent social space as more likely to be tied. The resulting posterior predictive distributions can provide a (model-based) lens into how well a network of one size compares to a network of a different size. Building on past work, we examine how the geometry of the latent space adjusts this lens and can further illuminate network differences in these forecasted distributions. We demonstrate how these prediction scores can be used to infer interesting behavioral patterns from an online social science experiment that examines group innovation in a competitive atmosphere.

EO137 Room R19 STATISTICS IN NEUROSCIENCE II Chair: Russell Shinohara

E0701: **tidyfun: A new framework for representing and working with function-valued data**

Presenter: **Jeff Goldsmith**, Columbia University, United States

Co-authors: Fabian Scheipl, Julia Wrobel

A new R package, tidyfun, is presented for representing and working with function-valued data that presents a unified interface for dealing with regularly or irregularly observed function-valued data. The package follows the tidyverse design philosophy of R packages and is aimed at lowering the barrier of entry for analysts in order to quickly and painlessly analyse and interact with functional data and, specifically, datasets that contain both scalar and functional data or multiple types of functional data, potentially measured over different domains. We discuss the available feature set as well as forthcoming extensions and show some application examples.

E0757: **Improving the replicability of spatial extent inference via effect size thresholding**

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. Here, we use 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, robust inference procedures must be used to ensure that the procedure maintains accurate error rates at an arbitrary p-value cluster forming threshold; for this reason, SEI via Gaussian Random Field approximations is not a valid inference procedure. Future work may investigate how effect size thresholding affects SEI power in small sample sizes and meta-analytic results.

E0841: **How to assess the neuroanatomical correspondence between brain imaging, gene expression and histological data**

Presenter: **Aaron Alexander-Bloch**, University of Pennsylvania, United States

Comparing the spatial or neuroanatomical pattern of different brain maps is, increasingly, a fundamental part of the experimental logic used in neuroimaging research. Three example comparisons help illustrate this neuroanatomical correspondence problem: 1) cortical folding versus cortical thickness; 2) myelination versus post-mortem gene expression; 3) neuronal cell density versus sensory multimodality. Each of these three examples of neuroanatomical correspondence has been the basis of strong biological inference. But in many studies, the neuroanatomical correspondence problem has been addressed simply by visual comparison or by spatial statistics whose assumptions are clearly violated. We recently proposed a so-called spin test, which generates a null distribution of correspondence by applying random rotations to spherical representations of cerebral the cortex. Dozens of studies have now adopted (and adapted) the spin test. The spin test has also been subject to legitimate statistical criticism, and alternative approaches to the neuroanatomical correspondence problem have been proposed. Unsurprisingly, the various approaches have different

strengths and weaknesses when applied to different kinds of datasets – including the three illustrative cases considered above – leading to the conclusion that the best approach to the neuroanatomical correspondence problem depends on the details of the experimental context.

E1127: **Understanding changes in brain topology and connectivity through single-subject ICA with empirical population priors**
Presenter: **Amanda Mejia**, Indiana University, United States

A primary objective in resting-state fMRI studies is the localization of resting-state networks (RSNs), regions of the brain that tend to act in a coordinated way, as well as the functional connectivity between them. These spatial and temporal properties of brain organization may be related to disease progression and treatment, development, and aging, making them of potential scientific and clinical value. A common tool to estimate RSNs is independent component analysis (ICA). However, due to high noise levels in fMRI, population average RSNs are often obtained and form the basis for estimating functional connectivity. Subject-level spatial RSN features are ignored, leading to potential bias in functional connectivity estimates and providing no information on differences in RSN topology. In hierarchical ICA, information shared across subjects is leveraged to improve subject-level estimation, but fitting such models is computationally intensive. We have proposed template ICA, a single-subject ICA model employing empirical population priors, which is computationally efficient and provides accurate subject-level estimates of RSNs and the functional connectivity between then. We will describe how template ICA allows for identification of unique subject-level spatial features, which, along with functional connectivity, can be used to understand better changes in the brain related to disease, aging, and other subject-specific factors.

EO185 Room R20 NOVEL APPROACHES TO CAUSAL INFERENCE Chair: Joseph Antonelli

E0214: **Improved inference for heterogeneous treatment effects using real-world data subject to hidden confounding**
Presenter: **Shu Yang**, North Carolina State University, United States

The heterogeneity of treatment effect (HTE) lies at the heart of precision medicine. Randomized clinical trials (RCTs) are gold-standard to estimate the HTE but are typically underpowered. While real-world data (RWD) have large predictive power but are often confounded due to lack of randomization of treatment. We show that the RWD, even subject to hidden confounding, may be used to empower RCTs in estimating the HTE using the confounding function. The confounding function summarizes the impact of unmeasured confounders on the difference in the potential outcome between the treated and untreated groups, accounting for the observed covariates, which is unidentifiable based only on the RWD. Coupling the RCT and RWD, we show that the HTE and confounding function are identifiable. We then derive the semiparametric efficient scores and integrative estimators of the HTE and confounding function. We clarify the conditions under which the integrative estimator of the HTE is strictly more efficient than the RCT estimator. As a by-product, our framework can be used to generalize the average treatment effects from the RCT to a target population without requiring an overlap covariate distribution assumption between the RCT and RWD. We illustrate the integrative estimators with a simulation study and an application.

E0253: **Facility profiling under competing risks using multivariate prognostic scores: Application to kidney transplant centers**
Presenter: **Youjin Lee**, University of Pennsylvania, United States
Co-authors: Douglas Schaubel

The performance of healthcare facilities (e.g., hospitals, transplant centers) is often evaluated through time-to-event outcomes. We consider the case where, for each subject, the failure event is due to one of several mutually exclusive causes (competing risks). Since the distribution of patient characteristics may differ greatly by center, some form of covariate adjustment is generally necessary in order for center-specific outcomes to be accurately compared (to each other or to an overall average). We propose methods for comparing facility-specific cumulative incidence functions (CIF) to an overall average. The methods involve directly standardizing each facility's non-parametric CIF through a weight function constructed from a multivariate prognostic score. We formally define the center-specific estimands with a causal interpretation, and establish conditions for their identification. We derive large-sample properties of the proposed estimators, and evaluate operating characteristics through simulation. We then apply the proposed methods to evaluate the center-specific pre-transplant mortality and transplant CIFs using data from the Scientific Registry of Transplant Recipients.

E0301: **Hospital quality risk standardization via approximate balancing weights**
Presenter: **Luke Keele**, University of Pennsylvania, United States

Comparing outcomes across hospitals, often to identify underperforming hospitals, is a critical task in health services research. However, naive comparisons of average outcomes, such as surgery complication rates, can be misleading because hospital case mixes differ — a hospital's overall complication rate may be lower due to more effective treatments or simply because the hospital serves a healthier population overall. We develop a method of "direct standardization" where we re-weight each hospital patient population to be representative of the overall population and then compare the weighted averages across hospitals. Adapting methods from survey sampling and causal inference, we find weights that directly control for imbalance between the hospital patient mix and the target population, even across many patient attributes. Critically, these balancing weights can also be tuned to preserve sample size for more precise estimates. We also derive principled measures of statistical precision and use outcome modeling and Bayesian shrinkage to increase precision and account for variation in hospital size. We demonstrate these methods using claims data from Pennsylvania, Florida, and New York, estimating standardized hospital complication rates for general surgery patients. We conclude with a discussion of how to detect low performing hospitals.

E0370: **More powerful tests of the composite null hypothesis arising in mediation analysis**
Presenter: **Caleb Miles**, Columbia University, United States
Co-authors: Antoine Chambaz

The indirect effect of an exposure on an outcome through an intermediate variable is identified by a product of regression coefficients under standard causal mediation assumptions and linear models for the outcome and intermediate variable. Thus, the null hypothesis of no indirect effect is a composite null hypothesis, as the null holds if either regression coefficient is zero. A consequence is that existing hypothesis tests are either severely underpowered near the origin (i.e., both coefficients being small with respect to standard errors) or invalid. We propose hypothesis tests that (i) preserve level alpha type 1 error, (ii) meaningfully improve power when both true underlying effects are small relative to sample size, and (iii) preserve power when at least one is not. One approach uses sparse linear programming to produce an approximately optimal test for a Bayes risk criterion. Another gives a closed-form test that is minimax optimal with respect to local power over the alternative parameter space.

EO736 Room R21 RECENT ADVANCES IN BIOSTATISTICS Chair: Reza Drikvandi

E0729: **Estimands, missing data, and sensitivity analysis**
Presenter: **Geert Molenberghs**, UHasselt, Belgium

Estimands, a crucial topic in clinical trials, are considered. A connection is made with the much older use in survey sampling theory. Using an example from surrogate marker evaluation, it is discussed where information comes from data, design, and assumptions. The latter may be unverifiable, hence the need to perform sensitivity analysis. The setting is then broadened to various forms of enrichment; that is, every situation where the model contains more aspect than the data can provide information about. Subsets of the enrichment class are: (a) coarsening; (b) augmentation. The focus is then placed on incomplete data for the rest of the presentation. A general framework for missing data is given, starting from Rubins seminal work. The defining and transforming role of the National Academy of Sciences report from 2010 about the Prevention and Treatment of Missing Data in Clinical Trials is evocated. It is argued that the role of the patient should not be forgotten, next to academe, regulators,

and industry. It is shown that for every MAR model, there is a family of MNAR models that exhibits the same fit to the data. Hence, one cannot show that MAR holds or not, solely depending on the data. The implications for standard and sensitivity analyses are discussed.

E1052: **Reference-based sensitivity analysis for time to event data**
Presenter: **James Carpenter**, London School of Hygiene and Tropical Medicine, United Kingdom
Co-authors: Andew Atkinson , Tim Clayton, Mike Kenward
Survival analysis typically assumes censoring at random, i.e. that, conditional on covariates in the model, the distribution of event times is the same, whether they are observed or censored. When trial patients who remain in follow up stay on their assigned treatment, then analysis under this assumption broadly addresses the de jure, or while on treatment strategy estimand. In such cases, we may well wish to explore the robustness of our inference to more pragmatic assumptions about the behaviour of patients post censoring. Such questions can be explored for trials with continuous outcome data using reference-based multiple imputations. This has two advantages: (a) it avoids the user specifying numerous parameters describing the distribution of patients' post-withdrawal data and (b) it is, to a good approximation, information anchored, so that the proportion of information lost due to missing data under the primary analysis is held constant across the sensitivity analyses. We bring this approach to the survival context, proposing a class of reference-based assumptions appropriate for survival data. We report a simulation study exploring the extent to which the multiple imputation estimator (using Rubin's variance formula) is information anchored in this setting and then illustrate the approach by re-analysing data from a randomized trial, which compared medical therapy with angioplasty for patients presenting with angina.

E0572: **Inference for model-agnostic variable importance**
Presenter: **Marco Carone**, University of Washington, United States
Co-authors: Brian Williamson, Peter Gilbert, Noah Simon
In many applications, it is of interest to assess the relative contribution of features (or subsets of features) toward the goal of predicting a response, that is, to gauge the variable importance of features. Most recent work on variable importance assessment has focused on describing the importance of features within the confines of a given prediction algorithm. However, such an assessment does not necessarily characterize the prediction potential of features and may provide a misleading reflection of the intrinsic value of these features. To address this limitation, we propose a general framework for nonparametric inference on interpretable algorithm-agnostic variable importance. We define variable importance as a population-level contrast between the oracle predictiveness of all available features versus all features except those under consideration. We then propose a nonparametric efficient estimation procedure that allows the construction of valid confidence intervals and tests, even when machine learning techniques are used.

E1161: **Permutation and Bayesian tests for random effects in mixed models**
Presenter: **Reza Drikvandi**, Durham University, United Kingdom
Random effects are used in mixed models to account for the unknown between-subject variability as well as the within-subject correlation in longitudinal, multilevel, clustered and other correlated data. Since random effects are latent and unobservable variables, it is challenging to decide which random effects to include or exclude from the model. In statistical language, this would be equivalent to testing whether or not the variance components of random effects are zero. However, test for zero variance components is a nonstandard testing problem because the null hypothesis in on the boundary of the parameter space and consequently the standard tests such as likelihood ratio, Wald and score tests may not be easily used. We introduce permutation tests and Bayesian tests for testing random effects which avoid the issues with the boundary of parameter space. The proposed methods are illustrated via simulations and a real data application.

EO135 Room R22 ADVANCES IN BAYESIAN METHODS AND APPLICATIONS Chair: David Rossell

E0864: **Quantifying heterogeneity in brain imaging data**
Presenter: **Donatello Telesca**, UCLA, United States
The application of functional data analysis to functional brain imaging has seen the development of a robust set of techniques for statistical estimation and inference in multivariate and highly structured settings. Beyond regression on the mean structure, we inquire how differential heterogeneity affects patterns of co-variation through a general regression framework involving both the mean function and the covariance operator. The inference is based on a shrinkage framework exploiting rank regularization in infinite factor models, which avoids ad-hock truncations typical of functional principal components representations. Our methodological contribution is illustrated in several case studies, including functional brain imaging of implicit learning and brain function during sleep. Time permitting, beyond variability attributable to observed covariates, we discuss the notion of latent functional features and their role in functional brain imaging.

E1007: **Bayesian loss function selection with the Hyvarinen score**
Presenter: **Jack Jewson**, Universitat Pompeu Fabra and Barcelona Graduate School of Economics, Spain
Co-authors: David Rossell, Piotr Zwiernik
General Bayesian updating provides a coherent procedure for updating beliefs about the minimiser of a loss function which need no longer index a probability density. Importantly this allows for Bayesian learning using algorithms as well as models. However, methods for selecting which amongst a series of loss functions is more appropriate for the data at hand are currently primitive. Loss functions need no longer define normalised probability densities and are thus no longer scale-invariant. As a result, standard Bayesian model selection tools do not apply. Instead, we appeal to the homogeneity property of the Hyvarinen score to select between loss functions in a scale-invariant manner. The chosen loss function can be interpreted as the pseudo-probability model that best captures the data generating process's relative probabilities. Doing so guarantees consistency, meaning we are still able to detect the data generating model if it is under consideration. In particular, we focus on examples from robust regression, where we are able to estimate the hyperparameter of Tukey's loss, and binary classification, comparing the SVM to logistic regression.

E0934: **A cosmological structure formation prior for dark matter searches**
Presenter: **Alex Geringer-Sameth**, Imperial College London, United Kingdom
Ultrafaint dwarf spheroidal galaxies, rich in dark matter, but hosting few visible stars, provide our most sensitive probe of the microscopic nature of dark matter, a central mystery in astronomy and particle physics. The small observational samples of stars make it difficult to infer how dark matter is distributed in these systems and non-informative priors are usually imposed to make progress. We propose to overturn this framework by using our best physical theories of how cosmic structures evolve over time and how galaxies form within them to derive new informative and realistic priors. Surprisingly, when applied to optical and gamma-ray observations, the new priors yield constraints on dark matter particle physics that are significantly weaker than under non-informative priors.

E1077: **Searching for dusty corners: Understanding the prediction of the cross-section of returns**
Presenter: **Carlos Carvalho**, The University of Texas at Austin, United States
Bayesian nonparametric regression models will be presented in order to predict equity returns from various characteristics. We will focus on model modifications that will incorporate economic information, time variability and explore ways to develop interpretable summaries of otherwise black-box strategies.

EO756 Room R23 ADVANCES IN BAYESIAN REGRESSION MODELING Chair: Andres Barrientos

E0233: **Bayesian hypothesis testing with highly unbalanced designs**
Presenter: **Victor Pena**, Baruch College, City University of New York, United States
The behavior of Bayes factors is explored in situations where there is a categorical predictor with some levels that are hard to observe. Some prior specifications exhibit pathological behavior which is reminiscent of what has become to be known as Lindley's paradox, with the key difference that, in this instance, it is proper priors that are problematic. We identify classes of prior specifications that are well-behaved and describe potential avenues for future work.

E1092: **Bayesian inferences on uncertain ranks and orderings: Application to ranking players and lineups**
Presenter: **Garritt Page**, Brigham Young University, United States
Co-authors: Andres Felipe Barrientos, David Dunson
It is common to be interested in rankings or order relationships among entities. In complex settings where one does not directly measure a univariate statistic upon which to base ranks, such inferences typically rely on statistical models having entity-specific parameters. These can be treated as random effects in hierarchical models characterizing variation among the entities. We are particularly motivated by the problem of ranking basketball players in terms of their contribution to team performance. Using data from the United States National Basketball Association (NBA), we find that many players have similar latent ability levels, making any single estimated ranking highly misleading. The current literature fails to provide summaries of order relationships that adequately account for such uncertainty. Motivated by this, we propose a strategy for characterizing uncertainty in inferences on order relationships among players and lineups. Our approach adapts to scenarios in which uncertainty in ordering is high by producing more conservative results that improve interpretability. This is achieved through a reward function within a decision-theoretic framework. We apply our approach to data from the 2009-10 NBA season.

E1174: **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 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 predictor's influences to vary flexibly over time. Smooth time-varying additive random effects are used to capture subject-specific heterogeneity. 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.

E1178: **Bayesian generalized linear models for correlated data with fewer latent variables**
Presenter: **Maryclare Griffin**, University of Massachusetts Amherst, United States
Many challenges arise when simulating from a Bayesian generalized linear model posterior distributions in practice, especially when the observed data is assumed to be dependent. We focus on two challenges that stem from the introduction of one or more auxiliary latent variables for each observation. First, several popular methods for simulating from Bayesian generalized linear model posterior distributions rely on the introducing of an auxiliary random variable for each observation. These methods can scale poorly when the number of observations is large because they require not only additional posterior draws but also repeated expensive matrix calculations. Second, many of the most useful approaches for introducing dependence in the observed data do so by introducing a latent random variable with a dense but computationally prior covariance matrix. However, the computational conveniences offered by the prior covariance matrix may be absent (or appear to be absent) from the posterior. We introduce methods for addressing these challenges that take advantage of simple reparameterizations of the problem, advances in posterior mode computation, and modern sampling methods.

EO680 Room R24 RECENT ADVANCEMENTS IN HIGH DIMENSIONAL STATISTICS Chair: Shubhadeep Chakraborty

E0315: **Nonparametric tests for independence and equality of distributions in high dimensions**
Presenter: **Shubhadeep Chakraborty**, University of Washington, Seattle, USA, United States
Co-authors: Xianyang Zhang
The paper presents new metrics to quantify and test for (i) the equality of distributions and (ii) the independence between two high-dimensional random vectors. We show that the energy distance based on the usual Euclidean distance cannot completely characterize the homogeneity of two high-dimensional distributions in the sense that it only detects the equality of means and the traces of covariance matrices in the high-dimensional set up. We propose a new class of metrics which inherits the desirable properties of the energy distance and maximum mean discrepancy/(generalized) distance covariance and the Hilbert-Schmidt Independence Criterion in the low-dimensional setting. The new class is capable of detecting the homogeneity of/completely characterizing independence between the low-dimensional marginal distributions in the high dimensional setup. We further propose t-tests based on the new metrics to perform high-dimensional two-sample testing/independence testing and study their asymptotic behavior under both high dimension low sample size (HDLSS) and high dimension medium sample size (HDMSS) setups. The computational complexity of the t-tests only grows linearly with the dimension and thus is scalable to very high dimensional data. We demonstrate the superior power behavior of the proposed tests for homogeneity of distributions and independence via both simulated and real datasets.

E0215: **High-dimensional change-point detection using generalized distance metrics**
Presenter: **Xianyang Zhang**, Texas A&M University, United States
Co-authors: Shubhadeep Chakraborty
Change-point detection has been a classical problem in statistics, finding applications in a wide variety of fields. A nonparametric change-point detection procedure is concerned with detecting abrupt distributional changes in the data generating distribution, rather than only changes in mean. We consider the problem of detecting an unknown number of change-points in an independent sequence of high-dimensional observations and testing 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 several other existing procedures is illustrated via both simulated and real datasets.

E0308: **Statistical inference for networks of high-dimensional point processes**
Presenter: **Ali Shojaie**, University of Washington, United States
Fueled in part by recent applications in neuroscience, high-dimensional Hawkes process has become a popular tool for modeling the network of interactions among multivariate point process data. While evaluating the uncertainty of the network estimates is critical in scientific applications, existing methodological and theoretical work has primarily focused on estimation. To bridge this gap, we develop a new statistical inference procedure for high-dimensional multivariate Hawkes processes. The key ingredient to this inference procedure is a new concentration inequality on the first- and second-order statistics for integrated stochastic processes, which summarizes the entire history of the process. Combining recent results on martingale central limit theory with the new concentration inequality, we can characterize the convergence rate of the test statistics. We

investigate finite sample properties of our inferential tools using extensive simulation studies and demonstrate their utility in an application to a neuron spike train data.

E0857: **Score matching for high-dimensional graphical models**
Presenter: **Mathias Drton**, Technical University of Munich, Germany

A common challenge in the estimation of parameters of multivariate probability density functions is the intractability of the normalizing constant. For continuous data, the score matching method provides a way to circumvent this issue and is particularly convenient for graphical modeling. We will present regularized score matching methods for high-dimensional and possibly non-Gaussian graphical models. In particular, we will discuss generalizations of score matching for observations that are non-negative or otherwise constrained in their support.

EO389 Room R25 BRANCHING PROCESSES: THEORY, COMPUTATION AND APPLICATIONS II Chair: Ines M del Puerto

E0770: **Branching processes under model misspecification**
Presenter: **Anand Vidyashankar**, George Mason University, United States

Inference for parameters of branching processes with and without immigration is well-understood. However, when the true probability model is misspecified, statistical estimators' behavior and the resulting inference concerning parameters are unknown. We address the estimator's asymptotic behavior when the offspring distribution is misspecified. Specifically, focusing on the offspring mean, we investigate the limiting distribution of the estimators when the true distribution belongs to a family of distributions belonging to the Kullback-Leibler class. Based on these results, we derive useful insights into conditional and marginal inference for these processes. Applications to robust inference are also provided.

E0859: **Modeling of metabolic mutation evolutionary processes using branching random walks**
Presenter: **Vladimir Kutsenko**, Lomonosov Moscow State University, Russia
Co-authors: Yulia Makarova, Elena Yarovaya

Recent researches demonstrate that the development of complex metabolic mechanisms in bacteria can be caused by a series of mutations that change the nutrition behavior. We study the bacteria population with a mutation leading to a complete change in nutrition behavior. We propose two models of such population evolution based on continuous-time branching random walks (BRWs) on the multidimensional lattices. The underlying random walk is assumed to be symmetric, homogeneous in space and irreducible. The first model based on BRW with one type of particles allows describing the evolution of mutant bacteria. At the initial time, the lattice is empty. We assume that immigration of particles occurs with a constant intensity at each lattice point. After immigration, a particle (a mutant bacterium) can split into two, die or jump to another lattice points. The second model based on the two-type BRW allows analyzing the distribution of both mutant and non-mutant bacteria at every lattice point. Initially at any lattice point may be only a finite number of first-type particles (non-mutant bacteria). In this model, a particle (a mutant or a non-mutant bacterium) can die or generate two particles of any type. The asymptotic behavior of the first moments of particles of both types at every lattice point was studied.

E0954: **Multi-type branching random walks on multidimensional lattices**
Presenter: **Yulia Makarova**, Lomonosov Moscow State University, Russia
Co-authors: Daria Balashova, Elena Yarovaya

The focus is on continuous-time multi-type branching random walks (BRWs), which may be described in terms of birth, death, and walking of particles of different types on multidimensional lattices. At the initial time moment, there is at least one particle of each type at every lattice point. For BRWs with sources of branching at each lattice point the reproduction law of particles is described by a multi-type branching process, which means that every particle can produce offsprings of each type by own branching mechanism. Moreover, particles can walk over the lattice. We assume that an underlying random walk for every type of particles is symmetric, homogeneous in space and irreducible. The main objects are subpopulations of particles generated by a single particle of each type at every lattice point and all over the lattice. The differential equations for generating functions for such subpopulations and their factorial moments are obtained. Based on such results, the solutions for the first moments of particle subpopulations are studied in detail. The application of such models for describing the spread of epidemics is discussed.

E1063: **Inference based on approximate Bayesian computation methods for X-linked two-sex branching processes**
Presenter: **Miguel Gonzalez Velasco**, University of Extremadura, Spain
Co-authors: Cristina Gutierrez Perez, Alicia Leon Naranjo, Rodrigo Martinez Quintana

The evolution of the number of individuals carrying the alleles, R and r, of a gene linked to X chromosome has been described using a multitype two-sex branching process. The R allele is considered dominant, and the r allele is supposed to be recessive and defective, responsible for a disorder. Hemophilia, red-green color blindness or the Duchenne and Becker's muscular dystrophies are examples of this kind of diseases. For this model, we investigate the estimation of its main parameters from a Bayesian standpoint. Concretely, we apply the Approximate Bayesian Computation (ABC) methodology to approximate its posterior distributions. The accuracy of the procedure is illustrated and discussed by way of a simulated example developed with R.

CI027 Room R02 APPLIED TIME SERIES Chair: Michael Owyang

C0207: **Binary conditional forecasts**
Presenter: **Michael McCracken**, Federal Reserve Bank of St. Louis, United States
Co-authors: Michael Owyang, Joseph McGillicuddy

While conditional forecasting has become prevalent both in the academic literature and in practice (e.g., bank stress testing, scenario forecasting), its applications typically focus on continuous variables. We merge elements from the literature on the construction and implementation of conditional forecasts with the literature on forecasting binary variables. We use the Qual-VAR, whose joint VAR-probit structure allows us to form conditional forecasts of the latent variable which can then be used to form probabilistic forecasts of the binary variable. We apply the model to forecasting recessions in real-time and investigate the role of monetary and oil shocks on the likelihood of two U.S. recessions.

C0310: **Forecasting low-frequency macroeconomic events with high frequency data**
Presenter: **Ana Galvao**, University of Warwick, United Kingdom
Co-authors: Michael Owyang

High-frequency financial and economic activity indicators are usually time aggregated before forecasts of low-frequency macroeconomic events, such as recessions, are computed. We propose a mixed-frequency modelling alternative that delivers high-frequency probability forecasts (including their confidence bands) for these low-frequency events. The new approach is compared with single-frequency alternatives using loss functions adequate to rare event forecasting. We provide evidence that: (i) weekly-sampled spread improves over monthly-sampled to predict NBER recessions, (ii) the predictive content of the spread and the Chicago Fed Financial Condition Index (NFCI) is supplementary to economic activity for one-year-ahead forecasts of contractions, and (iii) a weekly activity index can date the 2020 business cycle peak two months in advance using mixed-frequency filtering.

C0567: **Industrial connectedness and business cycle comovements**
Presenter: **Michael Owyang**, Federal Reserve Bank of St Louis, United States
Co-authors: Amy Guisinger, Daniel Soques

The effect of economic shocks on business cycles fluctuations at an industry level may vary across industries. For example, monetary shocks may have disparate effects, depending on the responsiveness of the particular industry to variations in the interest rate. Moreover, shocks that originate in a single industry may propagate elsewhere, either up or downstream in the production chain. Thus, more connected industries may be more vulnerable to industry-specific economic shocks. However, any model of industrial connectiveness must account for the fact that national shocks will drive much of the inter-industry correlation. In light of this, we develop a panel Markov-switching model for industry-level data that incorporates several features relevant for sub-national analysis. First, we model industry-level trends to differentiate between cyclical downturns and the secular decline in an industry. Second, we incorporate a national-level business cycle that industries may or may not attach to. Third, we model co-movement off of the national-level cycle as factors that affect clusters of industries.

CO111 Room R03 TIME SERIES ECONOMETRICS II Chair: Josu Arteche

C0363: **Frequency domain local bootstrap in long memory time series**
Presenter: **Josu Arteche**, University of the Basque Country UPV/EHU, Spain

Bootstrapping time series requires dealing with the serial dependence existing in the sample. In the time domain, this task has usually been accomplished by the sieve and the block bootstraps, whose purpose is to resample among approximately independent quantities finally. Frequency domain bootstrap techniques, basically based on transforming time dependence into heteroscedasticity, are less popular but they provide reliable approximations of the distribution of some statistics of weakly dependent series. However, its validity in long memory series has not been analysed yet. 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. The bound is used to justify the use of this strategy for some statistics such as the weighted 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.

C0565: **Modelling persistence change in fractionally integrated models**
Presenter: **Luis Filipe Martins**, ISCTE-IUL, Portugal
Co-authors: Josu Arteche

In recent years a vast literature documenting changes in the historical behaviour of economic and financial time series has been put forward. The popular parsimonious long-memory ARFIMA model describes both short and long memories simultaneously. There has been proposed parametric local stationary long-memory models. A new approach is proposed to model persistence change in fractionally integrated models. The model's statistical properties, estimation and inference are also studied. Some asymptotic properties of the estimators are derived, and an empirical application to the world inflation rates is considered.

C0769: **Robust quantile time series in financial time series models**
Presenter: **Valderio Anselmo Reisen**, DEST-CCE-UFES, Brazil
Co-authors: Pascal Bondon, Ian Danilevicz

The quantile regression and the m-regression methods which are widely used for time-independent data are invoked. We propose a robust quantile estimator for short and long memory time series, as frequently found in financial data. Asymptotic results of the estimator are established for Gaussian time series. Monte Carlo simulations illustrate the proposed methodology's performance under different scenarios of time series with additive outliers and asymmetric errors. As an application, the method is used to model the S&P 500 index. As an additional contribution, the methodology is introduced in mixed models with time series covariates. In this context, a real data set collected in the Greater Vitória area, Brazil, is analyzed to quantify the impact of the particular matter (PM2.5) levels on the health of children with asthma problems.

C1147: **US sea level data: Time trends and persistence**
Presenter: **Luis Alberiko Gil-Alana**, University of Navarra, Spain
Co-authors: Guglielmo Maria Caporale

US sea level data are analyzed by using long memory and fractional integration methods. All series appear to exhibit orders of integration in the range (0, 1), which implies long-range dependence; further, significant positive time trends are found in the case of 29 stations located on the East Coast and the Gulf of Mexico, and negative ones in the case 4 stations on the North West Coast, but none for the remaining 8 on the West Coast. The highest degree of persistence is found for the West Coast and the lowest for the East Coast.

CO233 Room R07 DEVELOPMENTS IN CRYPTOCURRENCY AND BLOCKCHAIN Chair: Marco Lorusso

C0275: **Conditional tail dependence in major cryptocurrency markets**
Presenter: **Pierangelo De Pace**, Pomona College, United States
Co-authors: Jayant Rao

A daily dataset of five major cryptocurrencies is used to empirically examine the conditional tail dependence of their price returns between April 2013 and March 2020. We do so by adopting a time-varying conditional copula modelling approach. The results are heterogeneous. We show that time variation in the tail dependence is generally pronounced in all pairs of cryptocurrencies. With some exceptions, tail dependences are usually low until mid-2018 and become very large in approximately the last two years of the sample.

C0346: **On the role of stablecoins in the cryptoassets pricing dynamics**
Presenter: **Ladislav Kristoufek**, Institute of Information Theory and Automation, Czech Academy of Sciences, Czech Republic

Stablecoins, usually represented by and associated with the dominant Tether (USDT) token, have evolved into an important clock-piece of the whole cryptoassets system. Their main objective is to enable easy transactions between cryptoasset exchanges with a stable exchange rate mostly through pegging to and being backed by a fiat currency or a physical asset. The backing itself has become a controversial topic for the most dominant stablecoin and its role in the 2017 skyrocketing cryptoassets prices has attracted a lot of speculations in the community. However, the research interest into the stablecoins and their role in the cryptoassets price dynamics has been rather scarce. Here we provide a detailed analysis of interactions and dynamics between the set of 28 stablecoins, Bitcoin as the most dominant cryptoasset, and altcoins to examine whether the stablecoins in fact induce the price movements or not. We provide evidence that the stablecoins mostly reflect an increasing demand for investing in cryptoassets rather than serve as a boosting mechanism for the periods of extreme appreciation. We further discuss some specificities of 2017, even though the general dynamic patterns remain very similar to the general behavior. Overall, we do not find support for various scandalous claims about the stablecoins.

C1035: **Cryptocurrency adoption and successful ICOs**
Presenter: **Luis Puch**, Universidad Complutense de Madrid and ICAE, Spain

Widespread use of Initial Coin Offerings (ICOs) to raise funds from investors is taking place. We test theory for the evolution of inflation and cryptocurrency holdings that support a successful ICO. In so doing, we construct a data set that contains information on prices and market capitalization for the top cryptocurrencies available in the market. Then we combine this information with various details on the corresponding *yellow papers* provided by the issuers. In particular, we focus on various distinguishing characteristics of adoption plans, and therefore, conditional on initial

tokens and initial prices, we analyze the evolution of inflation and market capitalization for utility cryptocurrencies at early stages. Our findings suggest that consistency between the number of cryptocurrency holdings that agents are willing to accept and the rate of fall in the exchange rate of the cryptocurrency with respect to legal tender are crucial to identify successful ICOs.

C1166: **Heterogeneous agents and cryptocurrency**
Presenter: **Marco Lorusso**, Newcastle University, United Kingdom
Co-authors: Francesco Ravazzolo, Stefano Grassi

The aim is to develop and estimate a Dynamic Stochastic General Equilibrium (DSGE) model with heterogeneous agents 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 the MitISEM. This econometric method is based on the importance sampling and avoids the Markov Chain Monte Carlo (MCMC) algorithm. There are several advantages associated with this method, e.g., parallelization and no autocorrelation of the draws. Results indicate that heterogeneity matters for the magnitude of crypto-specific shocks to the economy.

CO229 Room R08 INFLATION DYNAMICS AND COMMUNICATIONS Chair: Edward Knotek

C0797: **Inflation with Covid consumption baskets**
Presenter: **Alberto Cavallo**, Harvard University, United States

The Covid-19 Pandemic has led to changes in consumer expenditure patterns that can introduce significant bias in the measurement of CPI inflation. We use data collected from credit and debit transactions in the US to update the official basket weights. We find that the Covid inflation rate is higher for both headline and core CPIs, particularly for low-income households. We find similar results in 12 of 17 additional countries. The difference is growing over time as consumers spend relatively more on food and categories with rising inflation, and relatively less on transportation and categories experiencing significant deflation.

C0453: **Belief-dependent pricing decisions**
Presenter: **Rodrigo Lluberas**, Banco Central del Uruguay, Uruguay
Co-authors: Serafin Frache, Javier Turen

The effects of inflation and idiosyncratic cost expectations on firms' price-adjusting decisions are studied. We explore a novel monthly survey data on firms' expectations in Uruguay. Through the survey, we can directly assess price-adjustment decisions with firms' expectations while controlling for time and state-dependent factors. While inflation expectations do not play any role in our results, firms' beliefs about an expected increase in their overall costs matter as they positively affect the probability of adjusting prices. The evidence is consistent with the presence of forward-looking pricing at the firm level. The expectation channel is, however, heterogeneous across firms and operates with a delay. We show that the effect is driven exclusively by large firms. Being the beliefs about costs more volatile than inflation, the null reaction to this later expectation is in line with the theoretical predictions of Rationally-Inattentive price-setters.

C0430: **Average inflation targeting and household expectations**
Presenter: **Edward Knotek**, Federal Reserve Bank of Cleveland, United States
Co-authors: Olivier Coibion, Yuriy Gorodnichenko, Raphael Schoenle

Using a daily survey of U.S. households, we study how the Federal Reserve's announcement of its new strategy of average inflation targeting affected households expectations. Starting with the day of the announcement, there is a tiny uptick in the minority of households reporting that they had heard the news about monetary policy relative to before the announcement. Still, this effect fades within a few days. That hearing news about the announcement does not seem to have understood the announcement: they are no more likely to identify the Fed's new strategy than others correctly, nor are their expectations different. When we provide randomly selected households with pertinent information about average inflation targeting, their expectations still do not change differently than when households are provided with information about traditional inflation targeting.

C0442: **Communicating data uncertainty: Experimental evidence for U.K. GDP**
Presenter: **James Mitchell**, University of Warwick, United Kingdom
Co-authors: Ana Galvao, Johnny Runge

Many economic statistics, like GDP and inflation, are measured with error. But estimates are commonly communicated without any direct quantitative indication of their uncertainty. To assess if and how the public interprets and understands UK GDP data uncertainty, we conduct two sets of a randomized controlled online experiment, one at a time of growth and one during a recession. The surveys are designed to assess: (1) perceptions of the uncertainty in single-valued GDP growth numbers; (2) the public's interpretation and understanding of uncertainty information communicated in different formats; and (3) how communicating uncertainty affects trust in the data and the producer of these data. We find that the majority of the public understands that there is uncertainty inherent in GDP numbers, but communicating uncertainty information improves the public's understanding of why data revisions happen. It encourages them not to take GDP point estimates at face-value but does not decrease trust in the data. We find that it is especially helpful to communicate uncertainty information quantitatively using intervals, density strips and bell curves.

CC815 Room R06 CONTRIBUTIONS IN FINANCIAL MODELLING AND APPLICATIONS Chair: Cristina Amado

C0802: **Market regime detection via realized covariances: A comparison between unsupervised learning and nonlinear models**
Presenter: **Vito Ciciretti**, Independent, Germany
Co-authors: Andrea Bucci

There is broad empirical evidence of regime switching in financial markets. The transition between different market regimes is mirrored in correlation matrices, whose time-varying coefficients usually jump higher in stressed regimes, leading to the failure of common diversification methods. We aim to identify market regimes from correlation matrix features and detect transitions towards stressed regimes, hence improving tail-risk hedging. Starting from the time series of fractionally differentiated sentiment-like future values (such as gold, VIX, dollar index, etc.), we first build the realized correlation matrix for each period. Hence, we calculate several correlation features such as the distribution of the eigenvalues, the cophenetic index, the condition number. On these features adjusted to deal with collinearity, we apply an unsupervised learning methodology, the agglomerative hierarchical clustering, for labelling two latent market regimes, and compare its regime identification accuracy with a smooth transition autoregressive model applied on the covariance series. Finally, we fit a random forest classifier and evaluate its SHAP values to understand the most important correlation features when filtering market regimes. Our results show that the stressed regime is easier to be classified and that the cophenetic index and the percentage of variance explained by the eigenvalues above the Marchenko-Pastur upper bound are the most relevant explaining variables.

C0516: **Subjective shadow rates at the zero lower bound**
Presenter: **Ethan Struby**, Carleton College, United States
Co-authors: Michael Connolly

Shadow rates during the 2008-15 zero lower bound (ZLB) period are estimated using forward rates on US Treasuries and forecasts of short term interest rates from the Blue Chip Financial Survey. We estimate a suite of models with two- and three-factor structures, with and without forecast data, and alternatively assuming forecasts are full information rational expectations (FIRE) forecasts or distorted. Our estimates imply deeper easing (relative to previous estimates) by the Federal Reserve during the ZLB period relative to previous estimates. Still, only the distorted-forecast

model matches the Federal Reserve's real-time statements about the timing of increased policy rates prior to 2014. We revisit how monetary policy affected asset prices and risk premia during the crisis and find (1) all but one specification finds evidence of a structural break in the effects of monetary policy before and after the Great Recession (2) The model with distorted forecasts attributes more variance of yields to the expected path of short-term rates than models that assume FIRE and (3) term premia appear to have been affected through changes in duration risk or the slope of the yield curve rather than scarcity of medium-term assets.

C0306: **The anatomy of government bond yields synchronization in the Eurozone**

Presenter: **Claudio Barbieri**, Universita Cote d'Azur and Scuola Superiore Sant'Anna, France

Co-authors: Mauro Napoletano, Mattia Guerini

The synchronization of eurozone government bond yields at different maturities is investigated. To this scope, we combine principal component analysis and random matrix theory. We find that synchronization depends upon yields maturity. Short-term rates are not synchronized. For medium- and long-term yields, instead, synchronization has been high at the onset of the Euro experience, lapsed between the financial crisis and the end of the sovereign debt crisis and partially recovered after 2015. We show the existence of a duality between our results and portfolio theory, and we point toward divergence trades as a source of the self-sustained asynchronous dynamics. Our results envisage synchronization as a requirement for the smooth transmission of conventional monetary policy.

C1078: **2-Gaussian distribution modeling of financial data**

Presenter: **Cesar Berci**, FEEC - UNICAMP, Brazil

Co-authors: Celso Pascoli Bottura

The inadequacy of the normal distribution for representing the empirical probability distribution model of financial data, due to heavy tails and the leptokurtosis of the assets returns, among other reasons, is a challenging problem. The use of a novel stable distribution function that can model heavy tails and leptokurtosis is usual for probabilistic modeling of financial data. Some authors suggest that the α-stable is the stable distribution that has the best fit for financial data. However, this kind of distribution has no finite variance or higher moments or even an analytical closed-form, limiting its uses. For instance, it is inadequate for applications based on volatility analysis. An alternative, the 2-Gaussian distribution, a linear combination of Gaussians curves, inspired on Radial Basis Neural Networks, simpler than α-stable distribution and more flexible than a Normal distribution, was applied to model assets returns, reaching superior results than the others distributions, including the α-stable.

Authors Index

Aarts, E., 79
Abadi, A., 121
Abe, H., 134
Abeille, M., 114
Acal, C., 66
Acar, E., 75
Acosta, M., 39
Adatorwovor, R., 130
Afrifa-Yamoah, E., 110
Afyouni, S., 131
Agapiou, S., 21
Agapitos, O., 62
Agliardi, E., 136
Aguilera, A., 66
Ahn, J., 4
Akca, E., 120
Alba-Fernandez, V., 53
Albert Smet, J., 27
Alemdjrodo, K., 18
Alessi, L., 136
Alexander-Bloch, A., 145
Alexeev, V., 62
Alexopoulos, T., 136
Algaba, A., 102
Allard, A., 111
Allayioti, A., 82
Allison, J., 54, 114
Alonso, F., 56
Alonso-Pena, M., 108
Alsefri, M., 121
Altmeyer, R., 130
Alvarez-Liebana, J., 119
Alvarez-Perez, G., 119
Amado, C., 90
Ambrogi, F., 54
Ameijeiras-Alonso, J., 2
Amo-Salas, M., 119
Amris, K., 73
Anderson, C., 89
Ando, T., 59
Andreou, C., 101
Andriyana, Y., 107
Angelini, C., 57
Angelini, G., 104
Angulo, J., 8, 56
Ansell, J., 113
Antonelli, J., 77
Antoniano-Villalobos, I., 3
Apel, M., 39
Araki, Y., 107
Arashi, M., 53, 55
Arbel, J., 77, 109
Arcagni, A., 16
Archimbaud, A., 116
Arcos, A., 55, 58
Ardia, D., 102
Arendarczyk, M., 31
Aretz, K., 23
Argiento, R., 76
Argueta, J., 58
Arima, S., 16
Arismendi-Zambrano, J., 39
Ariza-Lopez, F., 53
Arjas, E., 133
Arnone, E., 42
Arslan, O., 53
Arteche, J., 150
Artemiou, A., 96, 115
Asai, M., 58, 59
Asar, Y., 55
Asdrubali, P., 81
Aslett, L., 65
Asta, D., 141
Atkinson, A., 147
Au, K., 65
Aue, A., 141
Austin, J., 84
Avella-Medina, M., 71

Babek, O., 20
Babic, S., 106
Babii, A., 74
Bacallado, S., 128
Bacri, T., 115
Bagchi, P., 86
Bagdonas, G., 32
Bagdonavicius, V., 8
Bagnato, L., 53, 84
Bai, J., 59
Bai, Y., 18
Baker, R., 6
Balabdaoui, F., 43
Baladandayuthapani, V., 132
Balashova, D., 149
Ballinari, D., 82
Bampinas, G., 80
Bandeira, A., 33
Bandi, F., 122
Bandyopadhyay, D., 21
Bandyopadhyay, S., 96
Banerjee, M., 43, 74
Banerjee, T., 48
Bantis, L., 37, 38
Bar, H., 30
Barbaglia, L., 102
Barbieri, C., 152
Bargagli Stoffi, F., 95
Bargigli, L., 38
Barltrop, C., 94
Barrera, A., 78
Barrett, J., 117
Barrientos, A., 21, 148
Bartolucci, F., 51
Barunik, J., 124
Basha, L., 83
Basir, T., 39
Basse, G., 86
Bassett, R., 141
Basu, S., 132
Battiston, S., 25
Bautista Barcena, M., 65
Bax, K., 11
Baxevani, A., 31, 101
Beaumont, J., 134
Bec, F., 123
Beck, B., 77
Beer, W., 66
Behrens, D., 66
Bekker, A., 53
Bellotti, A., 11
Beltran, D., 50
Benita, F., 17
Benvenuti, F., 62
Beraha, M., 35, 76
Beranger, B., 3
Berci, C., 152
Berentsen, G., 115
Bernardi, M., 57, 99
Bernardini, D., 11
Berrett, T., 71
Bersimi, E., 24
Bertarelli, G., 134
Berthet, Q., 31
Bertolacci, M., 34
Bertsch, C., 39
Beskos, A., 5
Besse, P., 89
Bevilacqua, M., 72, 124
Bhattacharjee, M., 74
Bhattacharya, A., 85
Bhattacharya, B., 48
Bhattacharya, R., 74
Bhattacharya, S., 140
Biagetti, M., 10
Bian, Y., 142
Bianchi, D., 57
Bianchi, M., 124
Bianco, N., 57
Bielak, L., 92
Billio, M., 25, 80
Billor, N., 14
Bissiri, P., 36, 113
Blanche, P., 117
Blanco, G., 95
Blix Grimaldi, M., 39
Bluteau, K., 102
Bochkina, N., 21
Bodnar, T., 105
Bodory, H., 95
Boegelsack, J., 46
Bogdan, M., 11
Bolin, D., 96
Bombelli, I., 64
Bondon, P., 150
Bongiorno, E., 119
Borisy, G., 97
Borrotti, M., 89
Borup, D., 91
Boudt, K., 102
Boulfani, F., 116
Bousselmi, B., 37
Bouzas, P., 93
Bowden, J., 87
Bradic, J., 115
Brakatsoulas, P., 125
Brave, S., 38
Brodie, R., 42
Bruce, S., 35, 86
Brunel, V., 31, 71
Brunotte, G., 115
Brzyski, D., 44
Buccheri, G., 79
Bucci, A., 151
Buccianti, A., 19
Bulatovic, N., 68
Bulger, D., 8, 58
Bulla, J., 115
Bunch, J., 45
Burrows, D., 55
Butters, A., 38
Butucea, C., 71

Caamano Carrillo, C., 72
Caballero-Aguila, R., 120
Cadonna, A., 111
Calabrese, R., 11, 12, 113
Calauzenes, C., 114
Caldara, D., 26, 137
Camehl, A., 122
Campbell, J., 38
Campbell, T., 101
Campos-Roca, Y., 90
Canale, A., 35
Cantelmo, A., 12
Cape, J., 33
Capitaine, L., 141
Caporale, G., 150
Caporin, M., 17, 24, 92, 122
Caprio, M., 99
Caraiani, P., 12
Caron, F., 52, 56
Carone, M., 147
Carpenter, J., 147
Carreau, J., 94
Carroll, R., 46
Carvalho, C., 147
Casarin, R., 25
Cascaldi-Garcia, D., 137
Casero-Alonso, V., 119
Castelletti, F., 85
Castelnuovo, E., 104
Castle, J., 1
Castro Cepero, L., 72
Castro Martin, L., 55
Castro, M., 73
Catalano, M., 118
Cavallaro, M., 63
Cavallo, A., 151
Cavicchioli, M., 58
Celov, D., 49
Cerovecki, C., 14
Cerqueti, R., 4
Cerreia-Vioglio, S., 123
Cervino, S., 66
Cevid, D., 74
Chadha, J., 69
Chakraborty, A., 142
Chakraborty, S., 77, 148
Chambaz, A., 146
Chambers, R., 134
Chandler, G., 30
Chandna, S., 4
Chang, J., 14
Chang, Y., 136
Characiejus, V., 14
Charemza, W., 49, 91
Chari, A., 103
Charlett, A., 77
Charpiat, G., 109
Chatterjee, S., 141

Chaturvedi, I., 16
Chaudhari, P., 29
Cheang, M., 34
Cheimarioti, A., 125
Chen, C., 14
Chen, J., 62, 124
Chen, K., 128
Chen, L., 45, 130
Chen, P., 60, 61
Chen, Y., 27, 33, 34
Cheng, B., 19
Cheng, W., 12
Chernozhukov, V., 60
Cheung, K., 19
Chevalier, C., 25
Chiaromonte, F., 14, 55
Chib, S., 21
Chodnicka - Jaworska, P., 83
Chowdhury, M., 36
Chretien, S., 17
Christensen, B., 92
Christensen, K., 62, 130
Chu, A., 59
Chung, H., 85
Chzhen, E., 89
Ciciretti, V., 151
Cinfrignini, A., 65
Ciolek, G., 130
Cirillo, P., 79
Cizikoviene, U., 121
Claeskens, G., 79, 115
Clark, A., 69
Clarke, D., 42
Clarte, G., 6
Clayton, T., 147
Cleanthous, G., 36
Clement, E., 5
Clements, A., 111
Clements, M., 124
Cobb, L., 46
Cobo, B., 67
Coibion, O., 151
Colubi, A., 27
Comunale, M., 26
Conde Llinares, S., 127
Connolly, M., 151
Conrad, C., 24, 125
Consoli, S., 92, 102
Constable, T., 97
Coolen, F., 65
Cordell, H., 87
Corradi, V., 1
Cortese, G., 46, 117
Costantini, M., 24, 80
Costola, M., 24, 25
Coull, B., 47, 97
Coullon, J., 56
Cousido Rocha, M., 66
Crainiceanu, C., 99
Craiu, R., 22
Cremona, M., 14
Crimaldi, I., 64
Cripps, E., 34
Cripps, S., 34
Croissant, L., 114
Crook, J., 11, 12, 125
Crujeiras, R., 2, 108
Csabai, I., 68
Cuba-Borda, P., 137
Cui, H., 16
Cui, Y., 37
Cumming, J., 110
Cuparic, M., 114
Curtis, A., 21
Czado, C., 11, 79, 113

D Ambrosio, A., 108
Dabo, S., 17
Dagdoug, M., 134
Dahl, D., 133
Dahlhaus, T., 104
Dai, X., 27
Dalayan, A., 31
Dambrosio, A., 84
Dang, H., 14
Dang, S., 97
Danielius, T., 55
Danielova Zaharieva, M., 90
Daniels, M., 144
Danilevicz, I., 150
Daouia, A., 72
Dare, W., 25
Dargel, L., 116
Darolles, S., 25
Dashti, M., 21
Datta, A., 132
Datta, S., 39
Dave, C., 103
Davenport, S., 71
Davies, D., 66
Davis, S., 49
Day, G., 63
de Amo, E., 32
de Angelis, D., 77
De Angelis, L., 137
De Block, A., 102
De Canditiis, D., 57
De Cian, E., 25
de Gunst, M., 76
De Iorio, M., 118
de la Calle-Arroyo, C., 118
De Luca, G., 124
de Luna, X., 43, 132
De Pace, P., 150
de Una-Alvarez, J., 9, 46
Deardon, R., 143
Degras, D., 71
Deistler, M., 111
del Barrio, E., 89
del Puerto, I., 135
Deldossi, L., 119
Deli, Y., 137
Delle Monache, D., 81
Delmarcelle, O., 102
Demirhan, H., 109
Dempsey, W., 128
Dendramis, Y., 125
Deresa, N., 9
Derumigny, A., 5, 29
Desai, N., 132
Descary, M., 93
Dessertaine, A., 134
Destercke, S., 65
Dette, H., 94, 142
Dewhirst, F., 97
Dhar, S., 3
di Lego, V., 66
Di Marzio, M., 2
Diaz, I., 88
Dietrich, M., 76
Dilts Stedman, K., 103
Ding, S., 133
Diop, A., 55
Djeundje, V., 125
Do, K., 87
Dobler, D., 76
Dobriban, E., 97
Dobronyi, C., 40
Doerre, A., 76
Dolgun, A., 109
Donayre, L., 38
Dondelinger, F., 67
Dong, Y., 111
Dorn, J., 99
Dou, X., 58
Draeger, L., 136
Drikvandi, R., 147
Drton, M., 149
Drutsa, A., 113
Dryden, I., 42
du Roy de Chaumaray, M., 84
Dubey, P., 46, 93, 141
Duetting, P., 114
Dufouil, C., 75
Dufour, A., 80
Dufour, J., 138
Duncan, A., 7, 65
Dunson, D., 33, 35, 148
Dupuy, J., 37
Durand, R., 125
Duranovic, A., 137
Durante, D., 22, 33
Durante, F., 5, 32, 59
Duren, Y., 30
Dutta, R., 63

Eaton, E., 44
Ebner, B., 53
Egorova, O., 63
Egozcue, J., 5, 19
Einmahl, J., 15
Elias, A., 28
Emad, N., 55
Emelyanov, G., 139
Emery, X., 36
Emiliozzi, S., 81
Emmanouil, S., 5
Engle, R., 125
Enikeeva, F., 31
Erdemlioglu, D., 17, 121
Ertefaie, A., 44
Esposito, M., 80
Esquivel, F., 56
Eustache, E., 134
Evangelaras, H., 109
Ewans, K., 116
Ezer, D., 127

Facevicova, K., 5
Faff, R., 124
Falconio, A., 111
Falk, M., 72
Fang, Q., 128
Fang, Y., 97
Farbmacher, H., 64
Faria-e-Castro, M., 137
Fasani, S., 12
Fasano, A., 21, 22
Fattore, M., 16
Fayaz, M., 121
Febrero-Bande, M., 27, 119
Fechteler, G., 37
Feifel, J., 76
Felici, G., 55
Feller, A., 86
Feng, Y., 20
Fensore, S., 2
Fernandez Iglesias, E., 120
Fernandez Sanchez, J., 5, 32
Fernandez, A., 66
Fernandez-Fontelo, A., 14
Ferrari, F., 99
Ferrari, P., 67
Ferraro, M., 64
Ferreira, J., 53
Ferreira, M., 100
Ferreira, T., 137
Ferri-Garcia, R., 55
Fiecas, M., 34, 85
Figa Talamanca, G., 111
Figini, S., 137
Figuerola-Ferretti Garrigues, I., 139
Filzmoser, P., 19
Finck, D., 83
Fine, J., 130, 131
Finkenstadt, B., 34
Fiocco, M., 76
Flores-Lagunes, A., 95
Flouri, T., 7
Fokianos, K., 116
Forastiere, L., 64, 95
Forbes, C., 117
Forbes, F., 77
Fosdick, B., 144
Fout, A., 145
Frache, S., 151
Francisco-Fernandez, M., 2
Franco Villoria, M., 47
Franzolini, B., 118
Frazier, D., 60
Freo, M., 69
Frias Bustamante, M., 119
Friedman, J., 44
Fries, S., 23
Fritsch, M., 110
Frost, F., 75
Fruehwirth-Schnatter, S., 110, 111
Frunza, M., 139
Fryzlewicz, P., 114
Fu, H., 138
Fuchs, S., 4
Fukasawa, M., 51
Funovits, B., 9
Furlanetto, F., 81
Furmanczyk, K., 11

Futschik, A., 63

Gaigall, D., 114
Gajardo, A., 93
Galvao, A., 103, 149, 151
Gamboa, F., 89
Gao, J., 67
Gao, X., 73
Gao, Y., 29
Garcia-Camacha Gutierrez, I., 118
Garcia-Diego, F., 55
Garcia-Jorcano, L., 137
Garcia-Portugues, E., 119
Gardlo, A., 5
Garibal, J., 24
Gasperoni, F., 117
Gatto, A., 59
Gaynanova, I., 85
Geenens, G., 4
Gegout-Petit, A., 84
Gendre, X., 116
Genest, C., 32
Genetay, E., 107
Geng, J., 88
Gennatas, E., 44
Georgiou, S., 109
Gerds, T., 88
Geringer-Sameth, A., 147
Gerotto, L., 103
Gersing, P., 111
Gerstenberg, J., 114
Ghaderinezhad, F., 7
Ghezzi, E., 89
Ghosh, J., 100
Ghosh, S., 36
Giacalone, M., 4
Giacometti, R., 11
Giangreco, A., 10
Giannone, D., 137
Gibberd, A., 17
Gieschen, A., 58, 113
Gijbels, I., 72
Gil-Alana, L., 150
Gil-Bermejo, C., 57
Gil-Leyva Villa, M., 35
Gilbert, P., 147
Gill, A., 109
Gilmour, S., 63, 89
Giordano, F., 117
Girard, S., 72
Giurcanu, M., 73
Gjika, E., 83
Glas, A., 82
Gloter, A., 5, 52
Gneyou, K., 106
Goessling, F., 90
Goga, C., 134
Goia, A., 119
Gokalp Yavuz, F., 53
Goldmann, L., 11
Goldsmith, J., 145
Golovkine, S., 107
Gong, G., 38
Goni, J., 44
Gonzalez Velasco, M., 135, 149
Gonzalez-Manteiga, W., 27, 119
Gonzalez-Rodriguez, G., 27, 120
Gorbach, T., 132
Gordaliza, P., 89
Gorfine, M., 108
Gorodnichenko, Y., 151
Goulet Coulombe, P., 91
Gourieroux, C., 40
Gozzi, C., 19
Graffelman, J., 5
Graham, M., 65
Grainger, J., 116
Granger, E., 75
Granziera, E., 104
Grassi, S., 79, 151
Grebe, M., 24
Greven, S., 14
Griffin, M., 148
Grigoriev, D., 10
Grimes, T., 133
Grith, M., 23
Gronwald, M., 124, 125
Gruber, K., 122
Grumiau, C., 22, 23
Guarracino, M., 64
Guay, A., 123
Guerini, M., 152
Guerra, M., 65
Guerrero, M., 28
Guglielmi, A., 35, 76
Guidolin, M., 24, 39
Guidotti, E., 52
Guillaumin, A., 52
Guindani, M., 1
Guisinger, A., 103, 149
Gunawan, D., 6
Guney, Y., 53
Gunter, U., 48
Guo, Z., 74
Gutauskaite, E., 32
Gutierrez Perez, C., 135, 149
Gutierrez, L., 35

Hacioglu, S., 26
Hadj-Amar, B., 34
Hafizi, R., 63
Hale, J., 64
Haliplii, R., 61
Hall, A., 1
Han, J., 121
Hanbali, H., 111
Haneuse, S., 47, 48, 143, 144
Hansen, J., 92
Hanson, M., 50
Hantzsche, A., 69
Hara, H., 57
Harezlak, J., 44
Harris, D., 67
Hartl, T., 49
Harvey, D., 60
Hasan, M., 110
Haskell, Z., 140
Hasse, J., 62
Haupt, H., 136
Hauser, D., 50
Hauzenberger, N., 122
Haziza, D., 134
He, K., 128
He, S., 82
He, T., 138
He, Y., 15
He, Z., 142
Heard, N., 67
Hebenstreit, D., 63
Hedt-Gauthier, B., 48
Heiner, M., 47
Heinisch, K., 82
Hejazi, N., 44
Helin, T., 21
Henckel, L., 132
Henderson, D., 16
Hendry, D., 1
Hennig, C., 113
Henninger, F., 14
Herbst, E., 40
Herculano, M., 50
Hermoso-Carazo, A., 120
Hickman, M., 77
Hiraki, K., 12
Hirata, W., 12
Hirukawa, M., 10
Hizmeri, R., 9, 61, 112
Hoermann, S., 14, 142
Hofert, M., 75
Hojda, P., 69
Hong, S., 68
Hong, Y., 57
Hoque, M., 75
Horiguchi, A., 37
Horii, S., 56, 57
Hornung, R., 108
Hossain, S., 78
Hou, Y., 15
House, T., 143
Howey, R., 87
Hron, K., 19, 20
Hu, G., 88
Hu, X., 19
Hu, Y., 144
Hu, Z., 18
Huang, C., 60
Huang, D., 130
Huang, W., 111
Huang, X., 73
Hubbard, R., 18
Huber, M., 64, 95
Huckstepp, R., 34
Hui, S., 16
Hull, I., 39
Humpherson, E., 127
Huser, R., 28
Huskova, M., 114
Hwang, Y., 120

Iannario, M., 15
Ibragimov, R., 81, 82
Ieva, F., 113
Ignatieva, K., 62
Ignazzi, C., 5
Iikubo, Y., 56
Imai, K., 98
Imaizumi, M., 2
Inghelbrecht, K., 102
Inoue, A., 49
Inoue, S., 59
Insolia, L., 55
Izzeldin, M., 9, 61, 112

Jackson Young, L., 103
Jackson, C., 117
Jahan-Parvar, M., 50
Jalasjoki, P., 104
Jammoul, F., 142
Jana, K., 107
Jansen, M., 115
Janson, L., 95
Jara, A., 72, 73
Javed, F., 106
Jaworski, P., 32, 83
Jenkins, P., 63
Jennison, C., 36
Jenq, R., 87
Jentsch, C., 78, 138
Jewson, J., 147
Ji, D., 97
Ji, T., 100
Jiang, B., 74
Jiang, F., 45
Jiang, H., 97, 131
Jiang, S., 87
Jiang, X., 7
Jiao, X., 7, 138
Jimenez, R., 28
Jimenez-Gamero, M., 53, 54
Jin, J., 20, 23
Jing, W., 56
Joe, H., 75
Johannsen, B., 40
Johndrow, J., 22
Johnson, D., 133
Jokubaitis, S., 49
Jonathan, P., 116
Josephs, N., 45
Josse, J., 31
Ju, N., 128
Jumah, A., 48
Jung, S., 93

Kaino, Y., 51
Kamatani, K., 5
Kapounek, S., 49
Karas, M., 99
Karavias, Y., 50
Karmakar, B., 98
Karmakar, S., 17
Karoui, A., 37
Karvanen, J., 132
Kashlak, A., 127
Kasprzak, M., 7
Kaszowska-Mojsa, J., 82
Kato, K., 8
Kato, S., 53
Kattuman, P., 81
Ke, T., 20
Keele, L., 146
Kellner, R., 11
Kelly, G., 54
Kenah, E., 43
Kenne Pagui, E., 117

Kennedy, E., 98
Kenney, A., 55
Kent, J., 116
Kenward, M., 147
Keogh, R., 75, 109
Keribin, C., 51
Kew, H., 67
Khan, M., 109
Khodakarim, S., 121
Khorrami Chokami, A., 72
Kieslich, P., 14
Kiley, M., 139
Kilinc, M., 123
Kim, B., 93
Kim, C., 136
Kim, H., 136
Kim, J., 101, 131, 143
Kim, K., 36
Kim, S., 81, 135
Kiran Chandra, N., 35
Kiriliouk, A., 94
Klaschka, J., 121
Klatt, M., 27
Kleen, O., 24
Klopp, O., 31
Klusowski, J., 29
Klutchnikoff, N., 107
Knaus, P., 111
Knight, K., 98
Knotek, E., 151
Knudson, A., 140
Kobayashi, T., 61
Koch, G., 73
Koh, A., 87
Koh, J., 94
Kohn, R., 6, 52
Kohns, D., 122
Kokonendji, C., 106
Kolaczyk, E., 45
Kolaiti, T., 136
Kolamunnage-Dona, R., 121
Kolokolov, A., 122
Koltchinskii, V., 141
Kolycheva, V., 10
Kondor, D., 68
Kong, L., 15
Konietschke, F., 73
Kontoghiorghes, A., 79
Konzou, E., 106
Koop, G., 123
Kordzakhia, N., 121
Kornak, J., 85
Korsaye, S., 79
Koskela, J., 63
Koslovsky, M., 87
Kostic, A., 114
Kottas, A., 47
Koudou, E., 106
Koutra, V., 89
Kovacevic, R., 66
Koval, B., 110
Kozina, A., 125
Kozubowski, T., 31, 140
Kratz, M., 140
Kreiss, A., 78
Kreuter, F., 14
Krishnamurthy, A., 46
Kristoufek, L., 68, 150
Krupskiy, P., 75
Kukacka, J., 125
Kuldyshev, O., 10
Kulik, A., 5
Kumbhakar, S., 40
Kunst, R., 48
Kurisu, D., 8
Kurle, J., 138
Kurtek, S., 42
Kuschinski, N., 72
Kutsenko, V., 149
Kyriacou, M., 60

Laber, E., 18
Laffers, L., 64, 95
Lahiri, S., 32
Lajaunie, Q., 40
Lam, C., 12
Lam-Weil, J., 71
Lambert, M., 25
Lan, Z., 21
Landsman, Z., 4
Lane, A., 144
Langen, H., 64
Langenus, G., 102
Lanza Queiroz, B., 66
Laoiza Maya, R., 60
Larriba, Y., 108
Lartigue, T., 67
Laurent, T., 116
Laurent-Bonneau, B., 71
Lawson, A., 143
Lazar, E., 82, 124, 138
Le Gouic, T., 89
Lederer, J., 30
Lee, C., 93
Lee, D., 89
Lee, E., 106
Lee, K., 97, 120
Lee, S., 35, 51, 98
Lee, Y., 98, 146
Leete, O., 18
Legramanti, S., 33
Leipus, R., 49, 121
Leng, C., 129
Leng, L., 12
Leng, X., 15
Lenoel, C., 69
Leon Naranjo, A., 149
Leon, L., 90
Leonelli, M., 29
Leonida, L., 10
Lespinasse, J., 75
Lessmann, K., 61
Levi, F., 34
Levin, K., 45
Levis, A., 143
Levuliene, R., 8
Lewandowski, M., 91
Lewin, A., 118
Ley, C., 2, 106
Leybourne, S., 60
Li, D., 93, 129
Li, F., 64, 98
Li, H., 99
Li, L., 97
Li, Q., 87, 131
Li, R., 7, 99
Li, T., 129
Li, W., 36
Li, X., 32
Li, Y., 23, 68
Li, Z., 7, 34, 35, 84, 144
Liang, F., 140
Liebl, D., 119
Lijoi, A., 21, 118
Lin, F., 131
Lin, J., 131
Lin, L., 45
Lin, Z., 84
Linares-Perez, J., 120
Lindberg, D., 144
Lindgren, F., 12
Lindgren, G., 4, 30
Lindholm, M., 105
Ling, W., 144
Linton, O., 68
Lipowski, C., 106
Liquet, B., 97
Litvinova, S., 82
Liu, A., 78
Liu, D., 10, 49
Liu, F., 68
Liu, J., 95
Liu, P., 18
Liu, S., 92, 97
Liu, Y., 131
Liu, Z., 140
Liverani, S., 56
Lluberas, R., 151
Lo, M., 106
Loecher, M., 108
Loeffler, M., 33
Loizeau, X., 45
Loperfido, N., 4
Lopes, M., 100
Lopez Oriona, A., 63
Lopez Pintado, S., 14, 27
Lopez-Fidalgo, J., 118
Lopez-Perez, A., 119
Lopez-Pintado, D., 27
Loria, F., 40, 137
Lorusso, M., 151
Loubes, J., 71, 89
Lu, H., 77
Lu, W., 12
Lu, Z., 60
Luati, A., 69
Lucidi, F., 61
Luger, R., 138
Lumbreras Sancho, S., 139
Lundblad, C., 103
Lundborg, A., 127
Luo, S., 20
Luta, G., 73
Lyall, J., 98
Lyziak, T., 92

Ma, L., 58
Ma, Y., 74
Maaitah, A., 80
Maathuis, M., 132
Macchiarelli, C., 69
Machalova, J., 19
Maddanu, F., 123
Madrid, A., 8
Magnusson, M., 133
Mahoney, M., 29
Maillet, B., 24
Maiti, T., 140
Makarova, S., 49
Makarova, Y., 149
Makimoto, N., 61
Makinen, T., 64
Makov, U., 4
Malinsky, D., 132
Mammen, E., 1
Mamonov, M., 10
Mandal, S., 36
Maneesoonthorn, W., 60
Manganelli, S., 111
Manipur, I., 64
Manisera, M., 42
Manolopoulou, I., 47
Manstavicius, M., 32
Mantoan, G., 103
Manzan, S., 102
Maraj, K., 101
Marbac, M., 84
Marchese, M., 9
Marin, J., 90
Mark Welch, J., 97
Marquis, B., 115
Marshall, A., 54
Marta, C., 66
Marta, T., 25
Martella, F., 64
Martin, G., 6, 60, 67
Martin-Barragan, B., 113
Martin-Martin, R., 118
Martinez Quintana, R., 149
Martino, L., 22
Martinoli, M., 78
Martins, L., 150
Martinussen, T., 88
Maruri, H., 63
Marushkevych, D., 130
Masak, T., 86
Masci, C., 113
Masciandaro, D., 39
Massacci, D., 10
Masuda, H., 5
Mateau, J., 119
Mateu, J., 8
Matsui, H., 106
Matsushima, T., 56
Mattera, R., 4
Matthes, C., 40
Mauras, S., 113
Maurer, H., 60
McCabe, B., 57
McClung, N., 103
McCracken, M., 103, 149
McCullagh, P., 53
McDonald, D., 98, 141
McGee, G., 47
McGillicuddy, J., 149
McIntyre, S., 123
McKinley, T., 143
Mealli, F., 64

Medialdea, A., 8
Medina-Olivares, V., 12
Medous, E., 134
Mei-Jie, Z., 18
Meilan-Vila, A., 2
Meintanis, S., 114
Mejia, A., 146
Mellina, S., 69
Mena, R., 35
Menafoglio, A., 19
Meng, X., 22
Menvouta, E., 23
Mercatanti, A., 64
Mercuri, L., 51
Merlevede, B., 17
Metodiev, M., 45
Miao, R., 27, 129
Michailidis, G., 31, 74
Michailidou, K., 34
Miescu, M., 81
Miglio, R., 54
Migliorati, M., 42
Mildiner Moraga, S., 79
Miles, C., 146
Milito, S., 117
Millis, B., 46
Milosevic, B., 53, 114
Minuesa Abril, C., 135
Miranda, E., 65
Miranda, M., 85
Mishra, T., 80
Mistry, M., 25
Misumi, T., 2
Mitchell, H., 54
Mitchell, J., 103, 123, 151
Mitra, N., 98
Modugno, M., 137
Moeller, J., 76
Moews, B., 58
Molenberghs, G., 146
Molinero, R., 25
Mollica, C., 16
Monarcha, G., 25
Monasterolo, I., 25, 137
Montagna, S., 77
Monteiro, A., 10
Montes, I., 65
Monti, A., 15
Montiel Olea, J., 40
Moon, H., 128
Moradi Rekabdarkolaee, H., 37
Morales Navarrete, D., 72
Morana, C., 137
Moreira, C., 46
Moreno, P., 66
Morita, H., 135
Morita, S., 134
Morris, J., 85, 132
Mosier, B., 38
Mostoufi, M., 22
Moulines, E., 31
Muehlmann, C., 5
Mueller, H., 46, 93, 100, 141
Mueller, M., 46
Mueller, P., 76, 133, 134
Mukherjee, A., 29
Mukherjee, D., 43
Mukherjee, G., 48
Mukherjee, K., 115
Mukherjee, S., 48, 67, 99
Mumtaz, H., 12
Muni Toke, I., 5
Munk, A., 27
Murtazashvili, I., 10
Musolesi, A., 136
Musolesi, M., 47
Musta, E., 76
Mustafin, I., 139
Muthukumar, R., 29
Myroshnychenko, S., 127

Nabi, R., 74
Nadarajah, K., 67
Naderi, M., 53
Nagl, M., 11
Nagy, S., 28
Nai Ruscone, M., 84, 113
Naik, C., 56
Naito, H., 57
Nakajima, J., 59
Nakhaeirad, N., 53
Nandi, S., 96
Napier, G., 89
Napoletano, M., 152
Naranjo Albarran, L., 90
Narisetty, N., 28
Nasini, S., 17
Nathoo, F., 45
Nava, C., 41
Naveau, P., 94
Ndaoud, M., 33
Neely, C., 121
Neethling, A., 53
Nettleton, D., 140
Nevo, D., 108
Newcombe, P., 117
Neykov, M., 42
Nezakati Rezazadeh, E., 115
Ng, S., 98
Ngatchou-Wandji, J., 54
Nguyen, H., 51, 106
Ni, Y., 77, 85
Nichols, T., 131
Nielsen, M., 102, 130
Niklasson, V., 105
Nipoti, B., 36
Nishino, H., 123
Nixon, M., 21
Nobili, A., 81
Nolte, I., 68
Nolte, S., 68
Nordhausen, K., 5, 55, 116
Nott, D., 6

Obradovic, M., 54
Oelrich, O., 133
Ogihara, T., 51
Oh, R., 4
Oja, H., 116
Ojimelukwe, A., 46
Old, O., 124
Olhede, S., 52
Oliveira, T., 59
Ollivier, Y., 109
OMalley, M., 96
Ombao, H., 28
Omori, Y., 59
Onrubia, J., 57
Oosterlee, C., 79
Opitz, T., 28, 94
Ortu, F., 123
Ossola, E., 136
Osuntuyi, A., 25
Otto, S., 71
Owyang, M., 103, 149
Ozenne, B., 88

Paccagnini, A., 26, 39, 40
Paczek, K., 44
Padilla, O., 141
Padoan, S., 3, 72
Paganoni, A., 113
Page, G., 76, 148
Paine, F., 50
Palacios-Rodriguez, F., 94
Palarea-Albaladejo, J., 19
Palau, S., 135
Palla, K., 56
Paloviita, M., 92, 104
Pan, J., 82
Panagiotidis, T., 80
PanahBehagh, B., 54
Panaretos, V., 93, 127, 142
Pandey, G., 81
Pandolfi, S., 51
Pandolfo, G., 108
Panero, F., 52
Panorska, A., 31, 140
Panzera, A., 2
Panzica, R., 136
Papadogeorgou, G., 98
Papantonis, I., 62
Papapanagiotou, G., 80
Papathomas, M., 56
Pardo-Fernandez, J., 78
Park, H., 135
Park, J., 35, 84, 85, 136
Park, K., 93
Park, S., 120
Parker, B., 89
Parla, F., 26
Parmeter, C., 10
Parsa, M., 8
Pascoli Bottura, C., 152
Patacca, M., 111
Paterlini, S., 11, 92
Patilea, V., 76, 107
Paul, S., 45
Pavlioglou, S., 22
Pavlu, I., 20
Pawlowsky-Glahn, V., 5
Pedio, M., 24, 80
Pedone, M., 110
Pedregal, D., 62
Pedroni, P., 102
Pellizzari, P., 103
Pena, J., 132
Pena, V., 148
Peng, K., 42
Penikas, H., 125
Pennino, M., 66
Pennoni, F., 51
Pensky, M., 33
Perera, I., 60
Peresetsky, A., 40
Perez Sanchez, C., 90
Perez-Veiga, N., 78
Perico Ortiz, H., 139
Pericoli, F., 81
Perles, A., 55
Perrakis, K., 67
Petereit, J., 140
Peters, J., 127
Petersen, A., 46
Petersen, I., 144
Peterson, C., 87
Petturiti, D., 65
Peyhardi, J., 90
Phillips, P., 60
Pietropaoli, A., 69
Pillai, N., 22
Pintus, P., 69
Pircalabelu, E., 96, 115
Pirino, D., 122
Pittavino, M., 67
Plagborg-Moller, M., 40
Plaksienko, A., 57
Platen, E., 60
Plummer, M., 33, 67
Podgorski, K., 4, 31
Podolskij, M., 130
Poignard, B., 58
Pokharel, G., 47
Polak, P., 105
Poli, F., 122
Polonik, W., 30
Poncela, P., 81
Ponnet, J., 23
Poon, A., 123
Porcu, E., 36
Porro, F., 46
Poskitt, D., 67
Poterie, A., 107
Powell, E., 66
Pozuelo Campos, S., 119
Prasad, A., 75
Prescott, T., 6
Preston, S., 42
Pretorius, C., 114
Prevosto, M., 30
Priebe, C., 33
Priftis, R., 50
Primiceri, G., 38
Proietti, T., 62, 124
Prokhorov, A., 10
Proust-Lima, C., 75
Prskawetz, A., 66
Pruenster, I., 21, 118
Pruneda, R., 118
Pua, A., 110
Puch, L., 150
Puech, P., 134
Puelz, D., 86
Punzo, A., 53, 84

Qi, Z., 128
Qian, K., 14

Qian, M., 19
Qiao, X., 14, 128
Qin, L., 30
Qiu, J., 27, 142
Qiu, Y., 44
Quaini, A., 79
Quinlan Binelli, J., 73
Quintana, F., 35, 76
Quiroz, M., 52
Qyrana, M., 137

Rabia, N., 17
Radojicic, U., 55, 116
Raftapostolos, A., 123
Rahman, M., 63
Ramirez Hassan, A., 60
Ramirez, S., 55
Ramos-Guajardo, A., 120
Ranalli, M., 64
Randolph, T., 44
Ranzato, G., 117
Rao, J., 150
Rapach, D., 91
Raubenheimer, L., 54
Ravazzolo, F., 59, 151
Raveendran, N., 8
Rebaudo, G., 21
Reich, B., 21
Reichold, K., 9
Reiczigel, J., 121
Reimherr, M., 119
Reisen, V., 150
Reiter, J., 21
Rejchel, W., 11
Remillard, B., 78
Ren, X., 60
Ren, Z., 32
Reno, R., 122
Restaino, M., 46, 54, 117
Reusens, P., 102
Rho, S., 101
Rholes, R., 104
Ricci, L., 106
Rice, G., 142
Richards, J., 28
Richardson, T., 132
Rigat, F., 100
Rigollet, P., 89
Rigon, T., 33
Risser, L., 89
Ritov, Y., 43
Riva, F., 25
Rivas-Lopez, M., 118
Rivera-Rodriguez, C., 48
Rivieccio, G., 124
Rizopoulos, D., 75
Rizzelli, S., 3
Robert, C., 6
Robin, G., 31
Robinson, P., 93
Rodrigues, J., 21
Rodriguez Gallego, A., 139
Rodriguez-Aragon, L., 118
Rodriguez-Diaz, J., 118
Rodriguez-Hernandez, M., 118
Rodriguez-Poo, J., 66, 136
Roesch, D., 11
Roldan, J., 66
Roman-Roman, P., 78
Romelli, D., 39
Romero, J., 8
Romo, J., 27
Rompolis, L., 62
Rosen, O., 34
Rosenbaum, P., 98
Rossell, D., 85, 147
Rossi, L., 12
Rossi, R., 81
Roszkowska, S., 69
Rotem, R., 47
Rousseau, J., 52, 56
Rouviere, L., 107
Roy, S., 17
Royer, J., 124
Rubarth, K., 73
Rubera, G., 39
Rubin, M., 79
Rubin, T., 127
Rubin-delanchy, P., 12, 33, 67
Rudin, C., 127
Rueda, C., 108
Rueda, M., 67
Ruisi, G., 40
Ruiz-Castro, J., 46, 66
Ruiz-Fuentes, N., 93
Ruiz-Gazen, A., 116, 134
Ruiz-Medina, M., 119
Runge, J., 151
Russo, M., 100
Ruzzi, D., 79
Rybinski, K., 49, 91
Rychlik, I., 4
Ryder, R., 6
Rysz, M., 81

Sadhanala, V., 141
Safikhani, A., 18
Sahamkhadam, M., 104
Sahin, O., 11, 113
Saint-Pierre, P., 134
Salehi, M., 54
Salish, N., 71
Salomone, R., 52
Salvati, N., 134
Samartsidis, P., 77
Samorodnitsky, G., 30
Samworth, R., 27
Sanchez Fuentes, A., 57
Sanchez-Romero, M., 66
Sanchis, L., 137
Sang, H., 88
Sangalli, L., 42
Sanna Passino, F., 67
Sant, J., 63
Santos, A., 10
Saracco, J., 97
Sarantsev, A., 101
Sarkar, A., 148
Sarkar, P., 32
Sarkar, S., 96
Sarno, L., 10
Satten, G., 144
Sauer, S., 48
Saumard, A., 107
Saumard, C., 107
Savy, N., 134
Sawadogo, A., 106
Scalera, P., 15
Scealy, J., 2
Schaubel, D., 146
Scheike, T., 54, 88
Scheipl, F., 145
Schirripa Spagnolo, F., 134
Schissler, A., 140
Schmidt-Hieber, J., 29
Schnurbus, J., 110
Schoenle, R., 151
Schreuder, N., 31, 89
Schult, C., 138
Schutte, E., 91
Schwartzman, A., 71
Schweinberger, M., 30, 86
Scicchitano, S., 10
Scotti, C., 26
Seal, S., 132
Seaman, S., 77
Seibold, H., 107
Sekhposyan, T., 104
Selk, L., 136
Semenov, A., 10, 81
Semmler, W., 60, 61
Sen, P., 73
Sengupta, D., 107
Sengupta, S., 145
Seo, W., 102
Seong, D., 102
Serafin, D., 46
Seregina, E., 104
Seri, R., 78
Seshadri, P., 65
Severino, F., 123
Severn, K., 42
Shah, R., 30, 127, 128
Shang, H., 71, 93
Shao, X., 93
Sharpnack, J., 141
Shchetinin, E., 40
Shen, S., 56
Sheng, X., 49
Shestopaloff, A., 22
Shevchenko, P., 121
Shi, C., 12
Shi, J., 7, 84
Shi, P., 37
Shi, S., 106
Shi, Y., 87
Shimai, Y., 61
Shin, M., 21
Shinohara, R., 131
Shioji, E., 135
Shojaie, A., 148
Shou, H., 99
Shpitser, I., 74, 132
Shushi, T., 4
Sibbertsen, P., 136
Siden, P., 52
Sigalla, S., 33
Sila, J., 68
Silvapulle, M., 60, 82
Silvennoinen, A., 1
Silvestrini, A., 64
Simicek, D., 20
Simon, N., 147
Simone II, C., 44
Simoni, A., 21
Simpson, S., 131
Sina, A., 80
Singh, R., 36
Sipek, A., 121
Sisson, S., 3
Sivula, T., 133
Sjolander, A., 43
Skorniakov, V., 121
Skouralis, A., 13
Skrobotov, A., 10, 81
Slavkovic, A., 21
Slavtchova-Bojkova, M., 135
Small, D., 98
Smedts, K., 111
Smeekes, S., 102
Smith, A., 22, 90, 145
Smith, L., 43
Smuts, M., 54, 114
So, H., 59
So, M., 59
Soberon, A., 136
Soegner, L., 9, 110, 111
Soffritti, G., 51
Sofronov, G., 8, 56, 58
Soh, C., 58
Sohn, M., 144
Son, J., 48
Song, R., 12
Soques, D., 149
Sorge, M., 103, 104
Sousa, R., 80
Sousedik, B., 46
Souto Arias, L., 79
Spano, D., 63
Spektor, S., 127
Spiewanowski, P., 91
Spindler, M., 64
Srakar, A., 6
Stanislawska, E., 92
Starr, J., 97
Stefanini, L., 65
Steger, J., 68
Stein, S., 129
Steinberger, L., 71
Steinmetz, J., 138
Stensrud, M., 108
Stephens, D., 134
Steven, R., 45
Stewart, J., 30, 86
Stingo, F., 110
Stival, M., 99
Stlianakis, N., 66
Stoehr, J., 6
Straczkiewicz, M., 99
Struby, E., 151
Stuart, M., 142
Stufken, J., 36
Stupfler, G., 3, 72
Stylianou, S., 109
Stypka, O., 9
Su, L., 98

Su, W., 29
Su, Z., 77
Suda, J., 69
Sudell, M., 121
Sun, B., 73
Sun, J., 57
Sun, W., 48, 96
Sun, Z., 16
Swartz, T., 42
Sykulski, A., 52, 96, 116
Szendrei, T., 122

Tahri, I., 61
Takeda, Y., 2
Talavera, O., 91
Tallarita, M., 118
Talska, R., 19, 20
Tan, Z., 73
Tang, M., 32, 33
Tao, J., 44
Tao, R., 47
Taufer, E., 11
Tavakoli, S., 127
Tawn, J., 28
Taylor, C., 2
Taylor, J., 96
Tchetgen Tchetgen, E., 87, 88
Tchouya, R., 17
Tebaldi, C., 123
Tegge, A., 100
Teh, Y., 56
Telesca, D., 147
Telschow, F., 71
Teodoro, M., 59
Tepegjozova, M., 79
Terada, Y., 2
Terasvirta, T., 1
Thaweethai, T., 143
Theisen, R., 29
Thiery, A., 65
Thomas-Agnan, C., 6
Thorsen, E., 105
Tibshirani, R., 31
Tille, Y., 134
Tiozzo Pezzoli, L., 92
Tokdar, S., 77
Tomarchio, S., 84
Tommasi, C., 119
Torabi, M., 75
Torrente Orihuela, A., 27
Torres Castro, I., 65
Torres, A., 119
Torres-Ruiz, F., 78
Torri, G., 11
Torsney, B., 36
Tortu, C., 64, 95
Tosetti, E., 92
Toulemonde, G., 94
Toulis, P., 86
Toure, A., 106
Tovey, H., 34
Trapani, L., 10
Trippa, L., 100
Trojak, M., 69
Trojani, F., 79
Trottier, C., 90
Trueck, S., 121
Trufin, J., 23
Trutschnig, W., 5
Tschernig, R., 49
Tse, Y., 111
Tsionas, M., 112
Tsybakov, A., 33
Tu, W., 97
Tuebbicke, S., 95
Turen, J., 151
Turgut, B., 69
Tuzcuoglu, K., 92, 104
Tzavalis, E., 62, 125

Ubeda Flores, M., 32
Uberti, P., 137
Uchida, M., 51
Ulrych, U., 105
Umbach, S., 91
Ungar, L., 44
Urbanek, J., 99
Urquhart, A., 68, 124
Usseglio-Carleve, A., 72
Uzeda, L., 104

Vaillancourt, J., 78
Valdes, G., 44
Vallee, A., 134
Valles Codina, O., 39
van Delft, A., 94
van den Boom, W., 118
van der Laan, M., 44
Van Keilegom, I., 8, 9, 76, 120
Van Oirbeek, R., 23
Vandekar, S., 145
Vandervorst, F., 23
Vansteelandt, S., 88
Vantaggi, B., 65
Varas, I., 73
Varotto, S., 80
Vasnev, A., 111
Vattay, G., 68
Veale, M., 127
Vega Carrasco, M., 47
Vehtari, A., 133
Veiga, H., 90
Veliov, V., 66
Veliyev, B., 62
Ventrucci, M., 47
Ventz, S., 100
Verdonck, T., 22, 23
Veredas, D., 106
Verhasselt, A., 107
Verschelde, M., 17
Vicini, A., 42
Vidyashankar, A., 149
Vieu, P., 119
Vilar, J., 63
Villani, M., 52, 133
Virbickaite, A., 90
Virta, J., 116
Visagie, J., 114
Vladimirova, M., 109
Vocalelli, G., 79
Vogelsang, T., 101
Voigt, S., 24
Vollmer, S., 7

Wadud, S., 125
Waernbaum, I., 132
Wagner, H., 110
Wagner, M., 9
Wallace, M., 43
Wallin, J., 31
Walsh, C., 78
Wan, R., 12
Wang, C., 143
Wang, D., 23, 141
Wang, G., 16
Wang, H., 101
Wang, J., 101
Wang, L., 44, 88
Wang, M., 20
Wang, P., 68
Wang, Q., 7, 15
Wang, S., 82, 138
Wang, T., 27
Wang, V., 100
Wang, W., 60
Wang, X., 20
Wang, Y., 30, 32, 35, 63
Ward, O., 90
Watanabe, T., 59
Webber, R., 56
Weber, E., 24, 49
Wein, A., 33
Westerlund, J., 50
Westwood, D., 66
Wheelock, D., 103
Widyastuti, E., 124
Wielaard, J., 5
Wikle, C., 52
Wilke, R., 106
Williams, J., 100
Williamson, B., 147
Wilson, S., 65
Winkelmann, L., 130
Winter, T., 55
Wirjanto, T., 142
Witzany, J., 125
Wlodarczyk, P., 82
Wodecki, A., 91
Wolinski, P., 109
Wong, R., 27, 128
Wood, A., 2, 117
Woodruff, D., 97
Woods, D., 63
Wozny, L., 91
Wrobel, J., 131, 145
Wu, E., 16
Wu, M., 144
Wu, W., 3, 37
Wu, Y., 145
Wu, Z., 19
Wylomanska, A., 101
Wynn, H., 63

Xia, D., 129
Xiao, G., 87
Xie, H., 16
Xing, F., 16
Xing, L., 44
Xing, X., 95
Xu, G., 81
Xu, J., 73, 145
Xu, M., 17
Xue, X., 138
Xue, Y., 88

Yadav, R., 28
Yamamoto, M., 2
Yang, F., 97
Yang, J., 3
Yang, P., 36
Yang, S., 78, 128, 146
Yang, X., 121
Yang, Y., 15, 71
Yang, Z., 7, 133
Yao, F., 84
Yao, W., 129
Yao, Y., 133
Yao, Z., 27
Yarovaya, E., 149
Yi, M., 116
Yin, X., 89, 128
Yiu, S., 98
Yoshida, N., 52
Young, K., 85
Yu, C., 142
Yu, D., 130
Yu, S., 68, 145
Yuan, Q., 37
Yuan, Y., 14
Yuasa, S., 135
Yukich, J., 28
Yurochkin, M., 89
Yuzbasi, B., 55

Zaliapin, I., 140
Zammit Mangion, A., 52
Zamojski, M., 23
Zarzo, M., 55
Zaytsev, A., 40
Zemel, Y., 27
Zenga, M., 46, 54
Zhan, X., 87, 97
Zhang, D., 40, 142
Zhang, E., 129
Zhang, J., 99
Zhang, L., 28, 44, 87
Zhang, N., 124
Zhang, X., 21, 27, 39, 93, 148
Zhang, Y., 74
Zhao, H., 97
Zhao, Q., 128
Zhao, S., 131
Zhao, X., 68
Zhao, Y., 18, 86, 142
Zhao, Z., 15, 95
Zheng, C., 18
Zheng, T., 90, 145
Zhilova, M., 141
Zhong, M., 26
Zhou, J., 45, 79, 115
Zhou, S., 88
Zhou, W., 44, 67, 86
Zhou, Y., 128
Zhou, Z., 3
Zhu, J., 20

Zhu, M., 75
Zhu, T., 133
Zhu, Y., 58, 130
Zhu, Z., 27
Zipunnikov, V., 19
Zoia, M., 41
Zou, H., 44
Zu, Y., 60
Zuccolotto, P., 42
Zuniga, F., 140
Zwiernik, P., 147

